bsa350

Logistic Regression

Logistic regression is a method for predicting the outcomes of 'either-or' trials. Either-or trials occur frequently in research. A person responds appropriately to a drug or does not; the dose of the drug may affect the outcome. A person may support a political party or not; the response may be related to their income. A person may have a heart attack in a 10-year period; the response may be related to age, weight, blood pressure, or cholesterol. We discuss basic ideas of modeling in the section titled 'Basic Ideas' and basic ideas of data analysis in the section titled 'Fundamental Data Analysis'. Subsequent sections examine model testing, variable selection, outliers and influential observations, methods for testing lack of fit, exact conditional inference, random effects, and Bayesian analysis.

Basic Ideas

A binary response is one with two outcomes. Denote these as $y = 1$ indicating 'success' and $y = 0$ indicating 'failure'. Let p denote the probability of getting the response $y = 1$, so

$$\Pr[y = 1] = p, \quad \Pr[y = 0] = 1 - p. \tag{1}$$

Let predictor variables such as age, weight, blood pressure, and cholesterol be denoted $x_1, x_2, \ldots, x_{k-1}$. A logistic regression model is a method for relating the probabilities to the predictor variables. Specifically, logistic regression is a **linear model** for the logarithm of the odds of success. The odds of success are $p/(1-p)$ and a linear model for the log odds is

bsa347

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}. \tag{2}$$

Here the β_js are unknown regression coefficients. For simplicity of notation, let the log odds be

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}. \tag{3}$$

It follows that

$$p = \frac{e^\eta}{1 + e^\eta}, \quad 1 - p = \frac{1}{1 + e^\eta}. \tag{4}$$

The transformation that changes η into p is known as the 'logistic' transformation, hence the name logistic regression. The transformation $\log[p/(1-p)] = \eta$ is known as the 'logit' transformation. Logit models are the same thing as logistic models. In fact, it is impossible to perform logistic regression without using both the logistic transformation and the logit transformation. Historically, 'logistic regression' described models for continuous predictor variables $x_1, x_2, \ldots, x_{k-1}$ such as age and weight, while 'logit models' were used for categorical predictor variables such as sex, race, and alternative medical treatments. Such distinctions are rarely made anymore.

Typical data consist of independent observations on, say, n individuals. The data are a collection $(y_i, x_{i1}, \ldots, x_{i,k-1})$ for $i = 1, \ldots, n$ with y_i being 1 or 0 for the ith individual and x_{ij} being the value of the jth predictor variable on the ith individual. For example, Christensen [3] uses an example from the Los Angeles Heart Study (LAHS) that involves $n = 200$ men: y_i is 1 if an individual had a coronary incident in the previous 10 years; $k - 1 = 6$ with $x_{i1} =$ age, $x_{i2} =$ systolic blood pressure, $x_{i3} =$ diastolic blood pressure, $x_{i4} =$ cholesterol, $x_{i5} =$ height, and $x_{i6} =$ weight for a particular individual. The specific logistic regression model is

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}. \tag{5}$$

Note that the model involves $k = 7$ unknown β_j parameters.

There are two ways of analyzing logistic regression models: frequentist methods and Bayesian methods. Historically, logistic regression was developed after standard regression (*see* **Multiple Linear Regression**) and has been taught to people who bsa424 already know standard regression. The methods of analysis are similar to those for standard regression.

Fundamental Data Analysis

Standard methods for frequentist analysis depend on the assumption that the sample size n is large relative to the number of β_j parameters in the model, that is, k. The usual results of the analysis are estimates of the β_js, say $\hat{\beta}_j$s, and standard errors for the $\hat{\beta}_j$s, say

$\text{SE}(\hat{\beta}_j)$s. In addition, a likelihood ratio test statistic, also known as a *deviance* (D), and *degrees of freedom* (df) for the deviance are given. The degrees of freedom for the deviance are $n - k$. (Some computer programs give alternative versions of the deviance that are less useful. This will be discussed later but can be identified by the fact that the degrees of freedom are less than $n - k$.)

The β_js are estimated by choosing values that maximize the *likelihood function*,

$$L(\beta_0, \ldots, \beta_{k-1}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}, \qquad (6)$$

wherein it is understood that the p_is depend on the β_js through the logistic regression model. Such bsa200 estimates are called **maximum likelihood estimates**. The deviance can be taken as

$$D = -2\log[L(\hat{\beta}_0, \ldots, \hat{\beta}_{k-1})]. \qquad (7)$$

For example, in the LAHS data

Variable	$\hat{\beta}_j$	$\text{SE}(\hat{\beta}_j)$
Intercept	−4.5173	7.451
x_1	0.04590	0.02344
x_2	0.00686	0.02013
x_3	−0.00694	0.03821
x_4	0.00631	0.00362
x_5	−0.07400	0.1058
x_6	0.02014	0.00984
$D = 134.9,$	$df = 193$	

Given this information, a number of results can be obtained. For example, β_6 is the regression coeffi- bsa127 cient for weight. A 95% **confidence interval** for β_6 has endpoints

$$\hat{\beta}_6 \pm 1.96\ \text{SE}(\hat{\beta}_6) \qquad (8)$$

and is (0.00085, 0.03943). The value 1.96 is the 97.5% point of a standard normal distribution. To test H_0: $\beta_6 = 0$, one looks at

$$\frac{\hat{\beta}_6 - 0}{\text{SE}(\hat{\beta}_6)} = \frac{0.02014 - 0}{0.00984} = 2.05 \qquad (9)$$

The P value for this test is .040, which is the probability that a standard normal random variable is greater than 2.05 or less than −2.05. To test $H_0 : \beta_6 = 0.01$, one looks at

$$\frac{\hat{\beta}_6 - 0.01}{\text{SE}(\hat{\beta}_6)} = \frac{0.02014 - 0.01}{0.00984} = 1.03 \qquad (10)$$

and obtains a P value by comparing 1.03 to a standard normal distribution. The use of the standard normal distribution is an approximation based on having n much larger than k.

Perhaps the most useful things to estimate in logistic regression are probabilities. The estimated log odds are

$$\log\left[\frac{\hat{p}}{1-\hat{p}}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6. \qquad (11)$$

For a 60-year-old man with blood pressure of 140 over 90, a cholesterol reading of 200, who is 69 inches tall and weighs 200 pounds, the estimated log odds of a coronary incident are

$$\log\left[\frac{\hat{p}}{1-\hat{p}}\right] = -4.5173 + .04590(60) + .00686(140) - .00694(90) + .00631(200) - 0.07400(69) + 0.02014(200) = -1.2435. \qquad (12)$$

The probability of a coronary incident is estimated as

$$\hat{p} = \frac{e^{-1.2435}}{1 + e^{-1.2435}} = .224. \qquad (13)$$

To see what the model says about the effect of age (x_1) and cholesterol (x_4), one might plot the estimated probability of a coronary incident as a function of age for people with blood pressures, height, and weight of $x_2 = 140$, $x_3 = 90$, $x_5 = 69$, $x_6 = 200$, and, say, both cholesterols $x_4 = 200$ and $x_4 = 300$. Unfortunately, while confidence intervals for this p can be computed without much difficulty, they are not readily available from many computer programs.

Testing Models

The deviance D and its degrees of freedom are useful for comparing alternative logistic regression models.

The actual number reported in the example, $D =$ 134.9 with 193df, is of little use by itself without

another model to compare it to. The value $D = 134.9$ is found by comparing the 7 variable model to a model with $n = 200$ parameters while we only have $n = 200$ observations. Clearly, 200 observations is not a large sample relative to a model with 200 parameters, hence large sample theory does not apply and we have no way to evaluate whether $D = 134.9$ is an appropriate number.

What we can do is compare the 6 predictor variable model (full model) with $k = 7$ to a smaller model (reduced model) involving, say, only age and weight,

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_{i1} + \beta_6 x_{i6} \qquad (14)$$

which has $k = 3$. Fitting this model gives

Variable	$\hat{\beta}_j$	$\mathrm{SE}(\hat{\beta}_j)$
Intercept	−7.513	1.706
x_1	0.06358	0.01963
x_6	0.01600	0.00794

$D = 138.8, \qquad df = 197.$

To test the adequacy of the reduced model compared to the full model compute the difference in the deviances, $D = 138.8 - 134.9 = 3.9$ and compare that to a chi-squared distribution with degrees of freedom determined by the difference in the deviance degrees of freedom, $197 - 193 = 4$. The probability that a $\chi^2(4)$ distribution is greater than 3.9 is .42, which is the P value for the test.

There is considerable flexibility in the model testing approach. Suppose we suspect that it is not the blood pressure readings that are important but rather the difference between the blood pressure readings, $(x_2 - x_3)$. We can construct a model that has $\beta_3 = -\beta_2$. If we incorporate this hypothesis into the full model, we get

$$\begin{aligned}
&\log\left[\frac{p_i}{1-p_i}\right] \\
&= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + (-\beta_2) x_{i3} + \beta_4 x_{i4} \\
&\quad + \beta_5 x_{i5} + \beta_6 x_{i6} \\
&= \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i2} - x_{i3}) + \beta_4 x_{i4} \\
&\quad + \beta_5 x_{i5} + \beta_6 x_{i6} \qquad (15)
\end{aligned}$$

which gives $D = 134.9$ on $df = 194$. This model is a special case of the full model, so a test of the models has

$$D = 134.9 - 134.9 = 0, \qquad (16)$$

with $df = 194 - 193 = 1$. The deviance difference is essentially 0, so the data are consistent with the reduced model.

This is a good time to discuss the alternate deviance computation. Logistic regression does not need to be performed on binary data. It can be performed on binomial data. Binomial data consist of the number of successes in a predetermined number of trials. The data would be $(N_i, y_i, x_{i1}, \ldots, x_{i,k-1})$ for $i = 1, \ldots, n$ where N_i is the number of trials in the ith case, y_i is the number of successes, and the predictor variables are as before. The logistic model is identical since it depends only on p_i, the probability of success in each trial, and the predictor variables. The likelihood and deviance require minor modifications.

For example, in our $k = 3$ model with age and weight, some computer programs will pool together all the people that have the same age and weight when computing the deviance. If there are 5 people with the same age-weight combination, they examine the number of coronary incidents out of this group of 5. Instead of having $n = 200$ cases, these 5 cases are treated as one case and n is reduced to $n - 4 = 196$. There can be several combinations of people with the same age and weight, thus reducing the effective number of cases even further. Call this new number of effective cases n'. The degrees of freedom for this deviance will be $n' - k$, which is different from the $n - k$ one gets using the original deviance computation. If the deviance degrees of freedom are something other than $n - k$, the computer program is using the alternative deviance computation.

Our original deviance compares the $k = 3$ model based on intercept, age, and weight to a model with $n = 200$ parameters and large sample theory does not apply. The alternative deviance compares the $k = 3$ model to a model with n' parameters, but in most cases, large sample theory will still not apply. (It should apply better. The whole point of the alternative computation is to make it apply better. But with binary data and continuous predictors, it rarely applies well enough to be useful.)

The problem with the alternative computation is that pooling cases together eliminates our ability to use the deviance for model comparisons. In comparing our $k = 3$ model with our $k = 7$ model,

it is likely that, with $n = 200$ men, some of them would have the same age and weight. However, it is very unlikely that these men would also have the same heights, blood pressures, and cholesterols. Thus, different people get pooled together in different models, and this is enough to invalidate model comparisons based on these deviances.

Variable Selection

Variable selection methods from standard regression such as forward selection, backwards elimination, and stepwise selection of variables carry over to logistic regression with almost no change. Extending the ideas of best subset selection from standard regression to logistic regression is more difficult.

One approach to best subset selection in logistic regression is to compare all possible models to the model

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 \qquad (17)$$

by means of score tests. The use of score tests makes this computationally feasible but trying to find a good model by comparing various models to a poor model is a bad idea. Typically, model (17) will fit the data poorly because it does not involve any of the predictor variables. A better idea is to compare the various candidate models to the model that contains all of the predictor variables. In our example, that involves comparing the $2^6 = 64$ possible models (every variable can be in or out of a possible model) to the full 6 variable ($k = 7$) model. Approximate methods for doing this are relatively easy (see [11] or [3, Section 4.4]), but are not commonly implemented in computer programs.

As in standard regression, if the predictor variables are highly correlated, it is reasonable to deal with the *collinearity* by performing *principal components regression* (*see* **Principal Component Analysis**). The principal components depend only on the predictor variables, not on y. In particular, the principal components do not depend on whether y is binary (as is the case here) or is a measurement (as in standard regression).

bsa501

Outliers and Influential Observations

In standard regression, one examines potential **outliers** in the dependent variable y and unusual combinations of the predictor variables $(x_1, \ldots, x_{k-1})$. Measures of influence combine information on the unusualness of y and $(x_1, \ldots, x_{k-1})$ into a single number.

bsa462

In binary logistic regression, there is really no such thing as an outlier in the ys. In binary logistic regression, y is either 0 or 1. Only values other than 0 or 1 could be considered outliers. Unusual values of y relate to what our model tells us about y. Young, fit men have heart attacks. They do not have many heart attacks, but some of them do have heart attacks. These are not outliers, they are to be expected. If we see a lot of young, fit men having heart attacks and our model does not explain it, we have a problem with the model (lack of fit), rather than outliers. Nonetheless, having a young fit man with a heart attack in our data would have a large influence on the overall nature of the fitted model. It is interesting to identify cases that have a large influence on different aspects of the fitted model.

Many of the influence measures from standard regression have analogues for logistic regression. Unusual combinations of predictor variables can be identified using a modification of the leverage. Analogues to Cook's distance (*see* **Multiple Linear Regression**) measure the influence (*see* **Influential Observations**) of an observation on estimated regression coefficients and on changes in the fitted log odds ($\hat{\eta}_i$s). Pregibon [12] first developed diagnostic measures for logistic regression. Johnson [8] pointed out that influence measures should depend on the aspects of the model that are important in the application.

bsa424

bsa301

Testing Lack of Fit

Lack of fit occurs when the model being used is inadequate to explain the data. The basic problem in testing lack of fit is to dream up a model that is more general than the one currently being used but that has a reasonable chance of fitting the data. Testing the current (reduced) model against the more general (full) model provides a test for lack of fit. This is a modeling idea, so the relevant issues are similar to those for standard regression.

One method of obtaining a more general model is to use a basis expansion. The idea of a basis expansion is that for some unknown continuous function of the predictor variables, say, $f(x_1, \ldots, x_{k-1})$, the model $\log[p/(1-p)] = f(x_1, \ldots, x_{k-1})$ should be appropriate. In turn, $f(x_1, \ldots, x_{k-1})$ can be approximated by linear combinations of known functions belonging to some collection of basis functions. The basis functions are then used as additional predictor variables in the logistic regression so that this more general model can approximate a wide variety of possible models. The most common method for doing this is simply to fit additional polynomial terms. In our example, we could incorporate additional terms for age squared (x_1^2), age cubed (x_1^3), height squared times weight cubed ($x_5^2 x_6^3$), etc. Alternatives to fitting additional polynomial terms would be to add trigonometric terms such as $\sin(2x_1)$ or $\cos(3x_4)$ to the model. (Typically, the predictor variables should be appropriately standardized before applying trigonometric functions.) Other options are to fit wavelets or even splines (*see* **Scatterplot Smoothers**). See [4, Chapter 7] for an additional discussion of these methods.

A problem with this approach is that the more general models quickly become unwieldy. For example, with the LAHS data, an absolute minimum for a basis expansion would be to add all the terms $x_{ik}^2, k = 1, \ldots, 6$ and all pairs $x_{ik}x_{ik'}$ for $k \neq k'$. Including all of the original variables in the model, this gives us a new model with $1 + 6 + 6 + 15 = 28$ parameters for only 200 observations. A model that includes all possible second-order polynomial terms involves $3^6 = 729$ parameters.

An alternative to using basis expansions is to partition the predictor variable data into subsets. For example, if the subsets constitute sets of predictor variables that are nearly identical, one could fit the original model but add a separate intercept effect for each near replicate group. Alternatively, one could simply fit the entire model on different subsets and see whether the model changes from subset to subset. If it changes, it suggests lack of fit in the original model. Christensen [5, Section 6.6] discusses these ideas for standard regression.

A commonly used partitioning method involves creating subsets of the data that have similar $\hat{p}_i$ values. This imposes a partition on the predictor variables. However, allowing the subsets to depend on the binary data (through the fitted values $\hat{p}_i$) causes problems in terms of finding an appropriate reference distribution for the difference in deviances test statistic. The usual χ^2 distribution does not apply for partitions selected using the binary data, see [7].

Landwehr, Pregibon, and Shoemaker [9] have discussed graphical methods for detecting lack of fit.

Exact Conditional Analysis

An alternative to the methods of analysis discussed above are methods based on exact conditional tests and their related confidence regions. These methods are mathematically correct, have the advantage of not depending on large sample approximations for their validity, and are similar in spirit to *Fisher's exact test* (*see* **Exact Methods for Categorical Data**) for **2 × 2 contingency tables**. Unfortunately, they are computationally intensive and require specialized software. From a technical perspective, they involve treating certain random quantities as fixed in order to perform the computations. Whether it is appropriate to fix (i.e., condition on) these quantities is an unresolved philosophical issue. See [1, Section 5] or, more recently, [10] for a discussion of this approach.

Random Effects

Mixed models are models in which some of the β_j coefficients are unobservable random variables rather than being unknown fixed coefficients. These random effects are useful in a variety of contexts. Suppose we are examining whether people are getting adequate pain relief. Further suppose we have some predictor variables such as age, sex, and income, say, x_1, x_2, x_3. The responses y are now 1 if pain relief is adequate and 0 if not, however, the data involve looking at individuals on multiple occasions. Suppose we have n individuals and each individual is evaluated for pain relief T times. The overall data are $y_{ij}, i = 1, \ldots, n;\ j = 1, \ldots, T$. The responses on one individual should be more closely related than responses on different individuals and that can be incorporated into the model by using a random effect. For example, we might have a model in which, given an effect for individual i, say, β_{i0}, the y_{ij}s are independent with probability determined by

$$\log\left[\frac{p_{ij}}{1-p_{ij}}\right] = \beta_{i0} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}. \tag{18}$$

However, $\beta_{10}, \ldots, \beta_{n0}$ are independent observations from a $N(\beta_0, \sigma^2)$ distribution. With this model, observations on different people are independent but observations on a given person i are dependent because they all depend on the outcome of the common random variable β_{i0}. Notice that the variables x_{ijs} are allowed to vary over the T times, even though in our description of the problem variables like age, sex, and income are unlikely to change substantially over the course of a study.

Such random effects models are examples of bsa252 **generalized linear mixed models**. These are much more difficult to analyze than standard logistic regression models unless you perform a Bayesian analysis, see the section titled 'Bayesian Analysis'.

Bayesian Analysis

bsa047 Bayesian analysis (*see* **Bayesian Statistics**) involves using the data to update the analyst's beliefs about the problem. It requires the analyst to provide a probability distribution that describes his knowledge/uncertainty about the problem prior to data collection. It then uses the likelihood function to update those views into a 'posterior' probability distribution.

Perhaps the biggest criticism of the Bayesian approach is that it is difficult and perhaps even inappropriate for the analyst to provide a prior probability distribution. It is difficult because it typically involves giving a prior distribution for the k regression parameters. The regression parameters are rather esoteric quantities and it is difficult to quantify knowledge about them. Tsutakawa and Lin [13] and Bedrick, Christensen, and Johnson [2] argued that it is more reasonable to specify prior distributions for the probabilities of success at various combinations of the predictor variables and to use these to induce a prior probability distribution on the regression coefficients. The idea that it may be inappropriate to specify a prior distribution is based on the fact that different analysts will have different information and thus can arrive at different results. Many practitioners of Bayesian analysis would argue that with sufficient data, different analysts will substantially agree in their results and with insufficient data it is appropriate that they should disagree.

One advantage of Bayesian analysis is that it does not rely on large sample approximations. It is based on exact distributional results, although in practice it relies on computers to approximate the exact distributions. It requires specialized software, but such software is freely available. Moreover, the Bayesian approach deals with more complicated models, such as random effects models (*see* **Random Effects in Multivariate Linear Models: Prediction**), with bsa497 no theoretical and minimal computational difficulty. See [6] for a discussion of Bayesian methods.

References

[1] Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science* **7**, 131–153.
[2] Bedrick, E.J., Christensen, R. & Johnson, W. (1997). Bayesian binomial regression, *The American Statistician* **51**, 211–218.
[3] Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, 2nd Edition, Springer-Verlag, New York.
[4] Christensen, R. (2001). *Advanced Linear Modeling*, 2nd Edition, Springer-Verlag, New York.
[5] Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, 3rd Edition, Springer-Verlag, New York.
[6] Congdon, P. (2003). *Applied Bayesian Modelling*, John Wiley and Sons, New York.
[7] Hosmer, D.W., Hosmer, T., le Cessie, S. & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model, *Statistics in Medicine* **16**, 965–980.
[8] Johnson, W. (1985). Influence measures for logistic regression: another point of view, *Biometrika* **72**, 59–65.
[9] Landwehr, J.M., Pregibon, D. & Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models, *Journal of the American Statistical Association* **79**, 61–71.
[10] Mehta, C.R., Patel, N.R. & Senchaudhuri, P. (2000). Efficient Monte Carlo methods for conditional logistic regression, *Journal of the American Statistical Association* **95**, 99–108.
[11] Nordberg, L. (1982). On variable selection in generalized linear and related regression models, *Communications in Statistics, A* **11**, 2427–2449.
[12] Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics* **9**, 705–724.
[13] Tsutakawa, R.K. & Lin, H.Y. (1986). Bayesian estimation of item response curves, *Psychometrika* **51**, 251–267.

Further Reading

Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*, John Wiley and Sons, New York.
McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman and Hall, London.

Tsiatis, A.A. (1980). A note on a goodness-of-fit test for the logistic regression model, *Biometrika* **67**, 250–251.

RONALD CHRISTENSEN

bsa350

8 Logistic Regression

Abstract: Logistic regression is a method for predicting the outcomes of 'either-or' trials. Either-or trials occur frequently in research. A person responds appropriately to a drug or does not; the dose of the drug may affect the outcome. A person may support a political party or not; the response may be related to their income. A person may have a heart attack in a 10-year period; the response may be related to age, weight, blood pressure, or cholesterol. We discuss basic ideas of modeling in the section titled 'Basic Ideas' and basic ideas of data analysis in the section titled 'Fundamental Data Analysis'. Subsequent sections examine model testing, variable selection, outliers and influential observations, methods for testing lack of fit, exact conditional inference, random effects, and Bayesian analysis.

Keywords: bayesian analysis; influential observations; lack of fit testing; logistic regression; mixed models; model testing; outliers; random effects; variable selection;

Author Contact Address: University of New Mexico, Albuquerque, New Mexico, USA