Class notes 8

Logistic regression analysis

Motivation

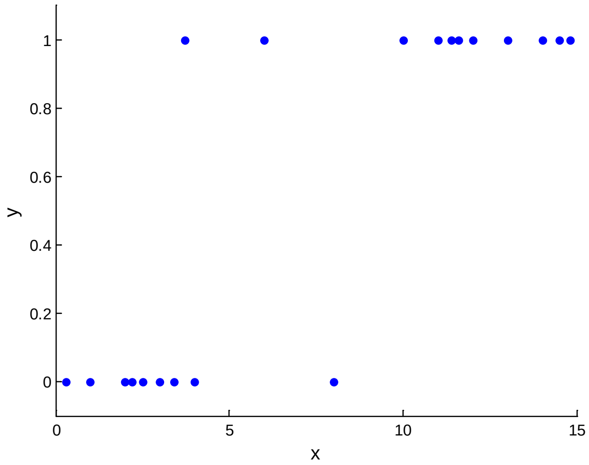
It is desirable to extend the linear regression modeling framework to allow the dependent variable to be categorical. In the following, we shall consider the simple case of Y being a variable with two levels. Suppose that Y is formulated as a binary variable, indicating either *success* ($Y = 1$) or *failure* ($Y = 0$) of some event occurring. An example could be the occurrence of a *heart attack*. The variable Y is then specified in the following way

$$Y = \begin{cases} 1 & \text{if the individual has suffered a heart attack} \\ 0 & \text{otherwise} \end{cases}$$

Another example could be the occurrence of being unemployed

$$Y = \begin{cases} 1 & \text{if the individual is unemployed} \\ 0 & \text{otherwise} \end{cases}$$

As for the linear regression model, it is of interest to relate Y to a set of independent variables (or predictors). For simplicity, consider the case of a single predictor. Let $\{(y_i, x_i)\}_{i=1}^{n}$ denote a set of n pairs of observations on X and Y. A scatter plot of the two variables is given below



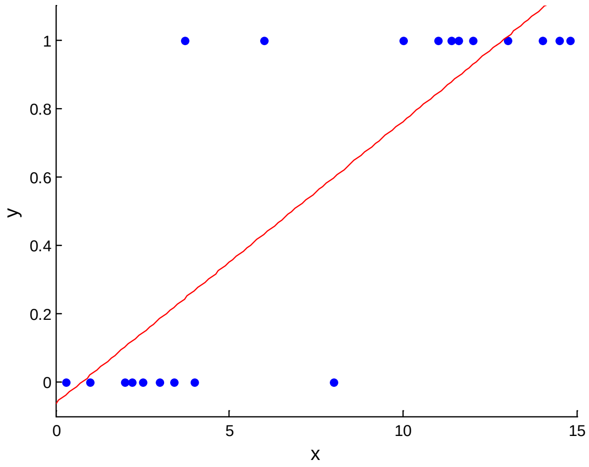
The plot indicates a positive relationship between X and Y. Clearly, larger values of X are associated with $Y = 1$, while smaller values are associated with $Y = 0$. Our primary interest is to model this association.

The linear probability model

Using the data, it is technically possible to fit a linear regression model of the form

$$Y = B_0 + B_1 X + \varepsilon$$

The result of doing so is illustrated in the following figure



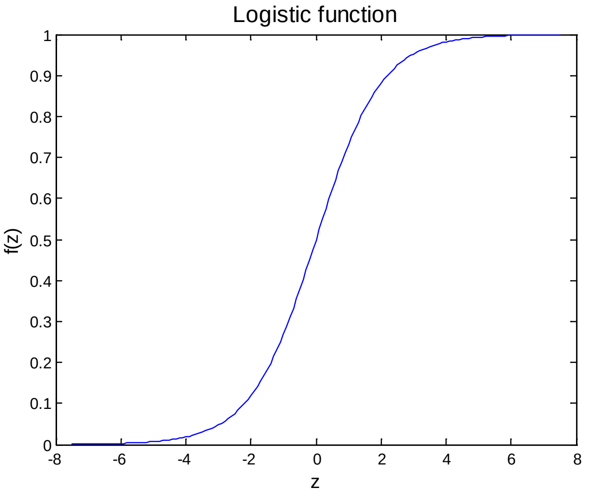
Since Y is binary, the expected value of Y (given some value of X) can be interpreted as the probability of the event occurring. More formally, this is stated as: $\mu_{Y|X} = P(Y = 1|X)$. The model is therefore known as the *Linear Probability Model* (LPM). However, fitting a linear model to the data may not be an optimal approach. First, an important shortcoming of the LPM is that the prediction of Y may lead to values larger than one or values less than zero. This is obviously inconsistent with the notion of probabilities. A second shortcoming is that the specified relationship is linear, which in some cases may not be realistic.

The logistic regression model

As an introduction to the logistic regression model, it is useful to define the logistic function given by

$$f(Z) = \frac{1}{1 + e^{-Z}}$$

where e denotes the exponential function and the mathematical variable Z is defined for all real numbers. The shape of the logistic function is illustrated below.



It now becomes a question of how to fit the logistic function to the data.

First, note that the exact shape of the logistic function can be manipulated by specifying a linear expression of the form

$$Z = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k$$

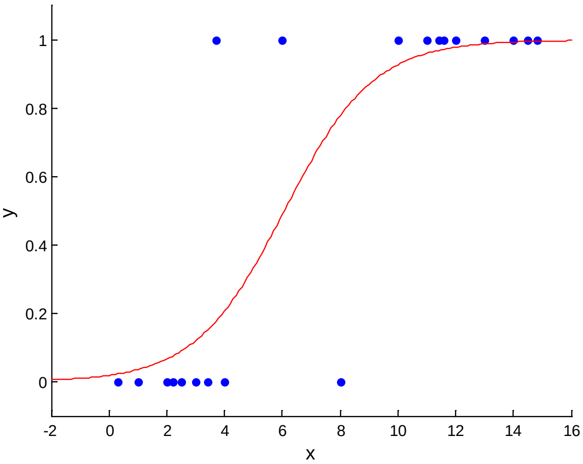
Using this expression, we obtain the general logistic regression model given by

$$\begin{aligned} P(Y = 1|X_1, \dots, X_k) &= \frac{1}{1 + e^{-(B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k)}} \\ &= \frac{e^{B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k}}{1 + e^{B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k}} \end{aligned}$$

In the case of a single predictor, the logistic regression model becomes

$$\begin{aligned} P(Y = 1|X) &= \frac{1}{1 + e^{-(B_0 + B_1X)}} \\ &= \frac{e^{B_0 + B_1X}}{1 + e^{B_0 + B_1X}} \end{aligned}$$

Fitting a logistic regression model to the data from before, we can clearly see that the logistic regression model represents a better fit to the data points as compared to the LPM



Interpretations

The logistic regression model can alternatively be written as a linear model in the form

$$\operatorname{logit}(P) = \log\left(\frac{P}{1 - P}\right) = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k$$

where we have used the simpler notation $P = P(Y = 1|X_1, \dots, X_k)$. The expression $P/(1 - P)$ is known as the *odds*. The odds is a quantity that can vary on a scale from zero to infinity. Note that the logistic regression model is a non-linear model. Hence, the effect of the predictors on the probability is not constant. To facilitate interpretation, it is useful to consider the model expressed in the terms of the odds

$$\frac{P}{1 - P} = e^{B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k}$$

From this expression, we see that the odds of $Y = 1$ when all the predictors are simultaneously zero is given by e^{B_0}. It is also of interest to evaluate the effect of the predictors on the odds. For instance, the effect on the odds of a one-unit increase in X_1, holding $X_2, \dots, X_k$ fixed (or constant), is given by e^{B_1}. For instance, consider the case when $B_1 = 0.15$, from which $e^{0.15} = 1.1618$. It then follows that a one-unit increase in X_1 translate into a 16.18% increase in the odds. More generally, when $B_j > 0$, an increase in X_j is associated with an increase in the odds. In the opposite case, when $B_j < 0$, an increase in X_j is associated with a decrease in the odds.

Estimation by maximum likelihood (ML)

For the logistic regression model, estimation works by maximizing a so-called *Likelihood function*, denoted $L(b_0, \dots, b_k; X_1, \dots, X_k)$, with respect to the model parameters. The exact formulation of the likelihood function is somewhat involved and is beyond the scope of these notes. For computational convenience, it is preferable to work with the natural logarithm of the likelihood function, here denoted $\log L(b_0, \dots, b_k; X_1, \dots, X_k)$, rather than the likelihood function itself. Since there are no closed form solutions for maximizing $\log L(b_0, \dots, b_k; X_1, \dots, X_k)$, the actual estimation is performed using a numerical *trial-and-error* procedure. The numerical algorithm searches over all possible values of $b_0, \dots, b_k$ until a maximum of the function is found. Many different methods for how to make the search algorithm perform as efficient as possible are available. If we are lucky, $\log L(b_0, \dots, b_k; X_1, \dots, X_k)$ is "well behaved" near its maximum such that $b_0, \dots, b_k$ can be determined in only a few trials (or iterations). Instead of the expression $\log L(b_0, \dots, b_k; X_1, \dots, X_k)$, we simplify the notation and use the shorter expression LL instead.

Model fit

To evaluate model fit, various statistical measures may be considered. Such measures include

- *Pseudo* R^2 (note that the usual R^2 used in linear analysis is not available for the logistic regression model. Therefore the term 'pseudo' is used): A range of different measures is available to the user. One of the more popular ones is the McFadden R^2, which is based on the log-likelihood function in the following way

  $$R^2 = 1 - \frac{LL_F}{LL_0} \quad \left(= 1 - \frac{-2LL_F}{-2LL_0}\right)$$

  In this expression, LL_F is the maximized log-likelihood obtained from estimating the full model (the model of interest), and LL_0 is the maximized log-likelihood obtained from estimating the so-called null-model. The null-model is a restricted model that does not contain any predictors (also referred to as the baseline model). The McFadden R^2 is conveniently bounded between 0 and 1, where 0 indicates no fit and 1 indicates perfect fit.

  Different software report different pseudo R^2 measures. For instance, SPSS reports the Cox & Snell R^2 and the Nagelkerke R^2. Unfortunately, the Cox & Snell R^2 possesses the undesirable property that its upper bound is less than one. The Nagelkerke R^2 is developed to correct this problem.

- *Testing for overall significance*: Just as we can apply the *F*-test to evaluate the simultaneously significance of all the predictors in a linear regression model, we can use a so-called *likelihood ratio* test to evaluate the same hypothesis for the logistic regression model.

In the general case of k predictors, the hypothesis for overall significance is

$$H_0 : B_1 = B_2 = \cdots = B_k = 0$$
$$H_A : \text{at least one } B_j \neq 0, \;\; j = 1,2,\dots,k$$

- *Hosmer-Lemeshow goodness-of-fit (GOF) test*: In short, the Hosmer-Lemeshow GOF test is based on a comparison between the probabilities predicted by the model and the corresponding probabilities observed in the sample. If this difference is not 'too' large, we conclude that the model represents an adequate fit. Performing the test, one is looking for a large *p*-value (larger than the significance level) to avoid rejecting the null-hypothesis of model fit. For a variety of reasons, the test has been subject to criticism. Note however that not many alternatives are available.

Hypothesis testing

Single restrictions

For the logistic regression model, testing the individual parameters can be performed using a variety of testing procedures, depending on the software.

Consider the two-tailed hypothesis

$$H_0 : B = B^*$$
$$H_A : B \neq B^*$$

The realized test statistic is

$$z = \frac{b - B^*}{SE(b)}$$

Under the null-hypothesis, the test statistic is asymptotically (when the sample size tend to infinity) standard normally distributed. If α denotes the significance level, the decision rule is to reject the null-hypothesis when the absolute value of z exceeds the critical value $z_{\alpha/2}$, where $z_{\alpha/2}$ is obtained from the normal table. The following table shows the critical value of α equals 0.01, 0.05 and 0.10:

α	$z_{\alpha/2}$ (Two-tailed test)
1 %	$z_{0.005} = 2.576$
5 %	$z_{0.025} = 1.960$
10 %	$z_{0.05} = 1.645$

Alternatively, the test can be performed using the *p*-value approach. The rejection rule is then

Reject the null hypothesis when the *p*-value $< \alpha$

As seen, testing the individual parameters in the context of a logistic regression model is slightly different from testing the individual parameters in a linear regression model. This difference arise because the logistic regression model is estimated using the ML approach rather than OLS. Testing procedures, such as the one presented above, rely on large sample theory (or asymptotic theory) and thereby make use of the normal distribution instead of the usual t-distribution.

Multiple restrictions

Recall that testing a set of linear restrictions when performing linear regression analysis involves estimating both a restricted model and an unrestricted model. The test is then performed applying the F-test.

In logistic regression analysis, a similar testing framework is available. In this case, the test is known as the *likelihood ratio* test, and is based on the log-likelihood function. Let LL_r denote the maximized log-likelihood obtained from estimating the restricted model, and let LL_{ur} denote the maximized log-likelihood obtained from estimating the unrestricted model.

The test statistic is obtained by

$$LR = -2 \cdot LL_r - (-2 \cdot LL_{ur})$$

Under the null-hypothesis, the LR-statistic is asymptotically χ^2-distributed with df equal to the number of restrictions being tested. The decision rule is to reject the null-hypothesis when the LR statistic exceeds the critical value $\chi^2_{\alpha,df}$.

Assumptions

For completeness, we briefly state the underlying assumptions of the logistic regression model:

- The data is taken from a random sample. That is, observations are independent.
- The model is correctly specified. This assumption states that the logit function is the appropriate link function for the response variable, and that the logit of the response variable is a linear combination of the predictors. Moreover, all relevant predictors are included are included in the specification.
- Errors are independently distributed (does not have to be normally distributed).
- There is no exact linear relationship among the predictor variables $X_1, \dots, X_k$ (no perfect multicollinearity).