Introduction to Logistic Regression

Ling-Chieh Kung[*]

December 30, 2014

When we use regression to study potential factors for an outcome, the data type of the dependent variable plays an important role. When the dependent variable is quantitative (e.g., income, heights, widths, weights, temperature, sales, yield rates), we use ordinary regression, the one we introduced in lectures. When it is binary (e.g., successful/failed, good/bad, dead/survived), we cannot use ordinary regression. Instead, we should use *logistic regression*. Below we will use one example to demonstrate how to do logistic regression with R.

Consider the data set contained in "glm_Eg.txt," which contains three columns and 45 rows. The 45 rows record relevant information of 45 persons who got trapped in a storm during a mountain hiking. Unfortunately, some of them died due to the storm. We are interested in predicting the survival probability of a person given her/his gender and age.[1] In the file, 0 means death and 1 means survival.

How to tackle this problem? Immediately we may want to run a linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where x_1 is one's age, x_2 is 0 if the person is a male or 1 if female, and $y = 1$ if the person is survived or 0 if dead. By running `lm()` (see the scripts contained in "glm_Eg.R"), one obtains the regression line

$$y = 1.066 - 0.013 x_1 - 0.319 x_2.$$

However, this is wrong! Suppose we are given a person's age as 50 and gender as male, we will predict that y equals $1.066 - 0.013 \times 50 = 0.416$. Obviously, y *cannot* be 0.416: It should be either 0 or 1! Even if we consider this fractional value as the probability for y to be 1, this is still problematic. For example, for a female who is 80 years old, the "probability" becomes $1.066 - 0.013 \times 80 - 0.319 = -0.293$, which is impossible! In general, using ordinary regression may get a predicted "probability" not within 0 and 1. This is why we should not do this.

The right way to do is to run a logistic regression. For a logistic regression, we hypothesize that independent variables x_is affect π, the probability for y to be 1, in the following form:[2]

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Given this functional form, the logistic regression model searches for coefficients to make the curve fit the given data points in the best way. While the details are far beyond the scope of this course, getting the estimated coefficients is easy. In R, all we need to do is to switch from

---

[*] Department of Information Management, National Taiwan University; lckung@ntu.edu.tw.

[1] The data set comes from the textbook *The Statistical Sleuth* by Ramsey and Schafer. The story has been modified.

[2] For our example, we only have two independent variables. In general we may have more.

`lm()` to `glm()` with an additional argument `binomial`.[3] Please see the script file for a concrete example. The estimated curve is

$$\log\left(\frac{\pi}{1-\pi}\right) = 3.23 - 0.078x_1 - 1.597x_2,$$

or equivalently,

$$\pi = \frac{\exp(3.23 - 0.078x_1 - 1.597x_2)}{1 + \exp(3.23 - 0.078x_1 - 1.597x_2)},$$

where $\exp(z)$ means e^z. It can be shown that π computed by the above equation will always lie in $[0, 1]$. There is thus no problem for interpreting π as a probability.

Let's depict the data and the fitted curves in Figure 1. The x-axis is for age and the y-axis is for the probability of survival. The blue and red dots are for males and females, respectively. The blue and red dashed lines are the regression lines given by ordinary regression, respectively. We can see that the probability of survival may indeed go beyond 1 or below 0. The blue and red solid curves are the regression curves given by logistic regression. They seem to be more reasonable predictions.

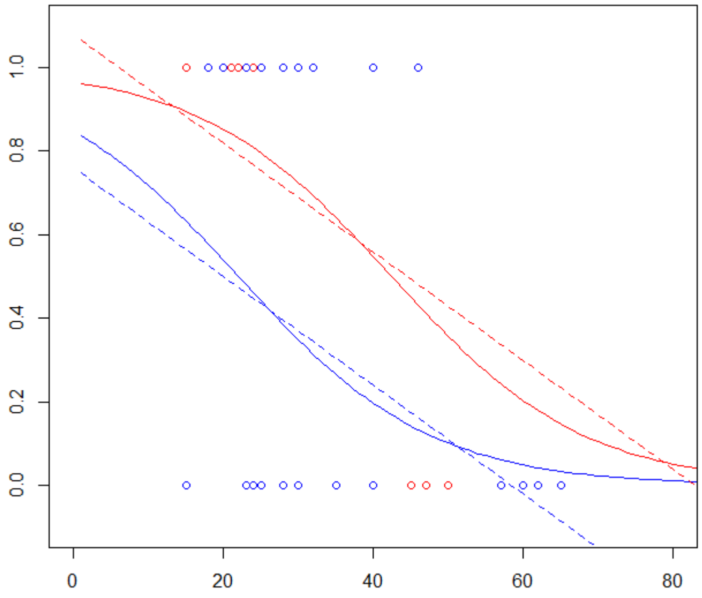

Figure 1: Regression curves for the examples

By reading the report for logistic regression, you may see that for each coefficient we still have an associated p-value. Just as for ordinary regression, we may use the p-value to determine whether the coefficient is significantly not 0 given a confidence level. In our example, both β_1 and β_2 are significantly not 0 at a 5% confidence level. Nevertheless, for logistic regression we do not have a simple measurement similar to R^2 for ordinary regression that can be used to evaluate the power of the model. In doing your final project, the best you may do is to give up interpreting your model in such a way.

---

[3] `lm` is the abbreviation of "linear model." `glm()` is the abbreviation of "generalized linear model."