Introduction to Logistic Regression

Ling-Chieh Kung[*]

December 21, 2015

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, we use ordinary regression, the one we introduced in lectures. When it is qualitative, we cannot use ordinary regression. Instead, we should use *logistic regression*. Below we will use one example to demonstrate how to do logistic regression for binary qualitative variables with R.

Note. See "survival.R" and "survival.txt" for R codes and data.

1 The story and data

Consider Table 1 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 *female* column, 0 means a man and 1 means a woman. In the *survival* column, 0 means death and 1 means survival.

Age	Gender	Survived	Age	Gender	Survived	Age	Gender	Survived
23	Male	No	23	Female	Yes	15	Male	No
40	Female	Yes	28	Male	Yes	50	Female	No
40	Male	Yes	15	Female	Yes	21	Female	Yes
30	Male	No	47	Female	No	25	Male	No
28	Male	No	57	Male	No	46	Male	Yes
40	Male	No	20	Female	Yes	32	Female	Yes
45	Female	No	18	Male	Yes	30	Male	No
62	Male	No	25	Male	No	25	Male	No
65	Male	No	60	Male	No	25	Male	No
45	Female	No	25	Male	Yes	25	Male	No
25	Female	No	20	Male	Yes	30	Male	No
28	Male	Yes	32	Male	Yes	35	Male	No
28	Male	No	32	Female	Yes	23	Male	Yes
23	Male	No	24	Female	Yes	24	Male	No
22	Female	Yes	30	Male	Yes	25	Female	Yes

Table 1: The survival data set

[*] 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 Descriptive statistics

As always, we start from descriptive statistics. We see that the overall survival probability is $\frac{20}{45} = 44.4\%$. Moreover, survival or not seems to be affected by gender:

Group	Survivals	Group size	Survival probability
Male	10	30	33.3%
Female	10	15	66.7%

By grouping people into age classes, it seems to us that survival or not is also affected by age:

Age class	Survivals	Group size	Survival probability
$[10, 20)$	2	3	66.7%
$[21, 30)$	11	22	50.0%
$[31, 40)$	4	8	50.0%
$[41, 50)$	3	7	42.9%
$[51, 60)$	0	2	0.0%
$[61, 70)$	0	3	0.0%

These findings are good, but they are just some descriptions of our sample. May we do better than just descriptive statistics? May we make some inferences about the population? Finally, may we predict one's survival probability given one's age and gender?

3 Why ordinary regression does not work?

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

$$\mathit{survival}_i = \beta_0 + \beta_1 \mathit{age}_i + \beta_2 \mathit{female}_i + \epsilon_i.$$

to our sample data. By running

```
d <- read.table("survival.txt", header = TRUE)
fitWrong <- lm(d$survival ~ d$age + d$female)
summary(fitWrong)
```

one obtains the regression line

$$\mathit{survival} = 0.746 - 0.013\mathit{age} + 0.319\mathit{female}.$$

Figure 1 illustrate the regression lines obtained from the ordinary regression fitting. Though $R^2 = 0.1642$ is low, both variables are significant. The result seems to be reasonable: Being younger or being a woman makes the predicted value of *survival* higher. This fits our observation in the previous section.

However, this is wrong! Consider an 80-year-old man. For him, the predicted survival "probability" becomes $0.746 - 0.013 \times 80 = -0.294$, which is impossible. In general, it is very easy for an ordinary regression model to generate predicted "probability" not within 0 and 1. This is why ordinary regression is problematic when the dependent variable is binary.

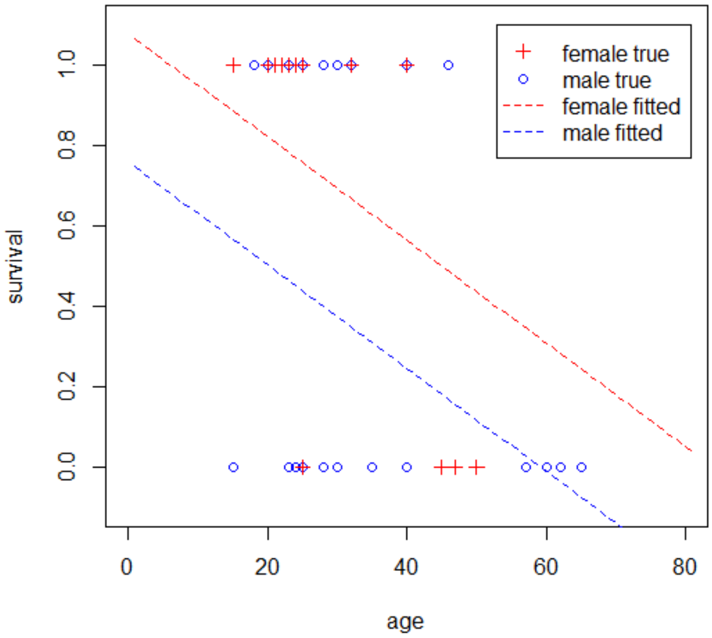

Figure 1: Ordinary regression fitting

4 Logistic regression

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

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

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 `lm()` to `glm()` with an additional argument `binomial`:[3]

```
fitRight <- glm(d$survival ~ d$age + d$female, binomial)
summary(fitRight)
```

By executing the above statements, we will get a regression report. While some information is new, the following part is familiar to us. We then understand that both variables are significant:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.63312    1.11018   1.471   0.1413
d$age       -0.07820    0.03728  -2.097   0.0359 *
d$female     1.59729    0.75547   2.114   0.0345 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

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

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

According to the regression report, the estimated curve is

$$\log\left(\frac{\pi}{1-\pi}\right) = 1.633 - 0.078\,age + 1.597\,female, \tag{1}$$

or equivalently,

$$\pi = \frac{\exp(1.633 - 0.078\,age + 1.597\,female)}{1 + \exp(1.633 - 0.078\,age + 1.597\,female)},$$

where $\exp(z)$ means e^z for all $z \in \mathbb{R}$. Obviously, π computed by the above equation will always lie in $[0, 1]$. There is thus no problem for interpreting π as a probability.

Figure 2 depicts the regression curves obtained by logistic regression fitting. Besides observing that they are indeed bounded by 0 and 1, they can be used directly for prediction. For example, for the man at 80, we have

$$\pi = \frac{\exp(1.633 - 0.078 \times 80)}{1 + \exp(1.633 - 0.078 \times 80)} = 0.0097,$$

which is no longer the unreasonable -0.294 obtained by ordinary regression. As another example, for a woman at 60, we have

$$\pi = \frac{\exp(1.633 - 0.078 \times 60 + 1.597)}{1 + \exp(1.633 - 0.078 \times 60 + 1.597)} = 0.1882.$$

These probabilities can be verified by investigating the two curves in Figure 2. Figure 3 provides a comparison.

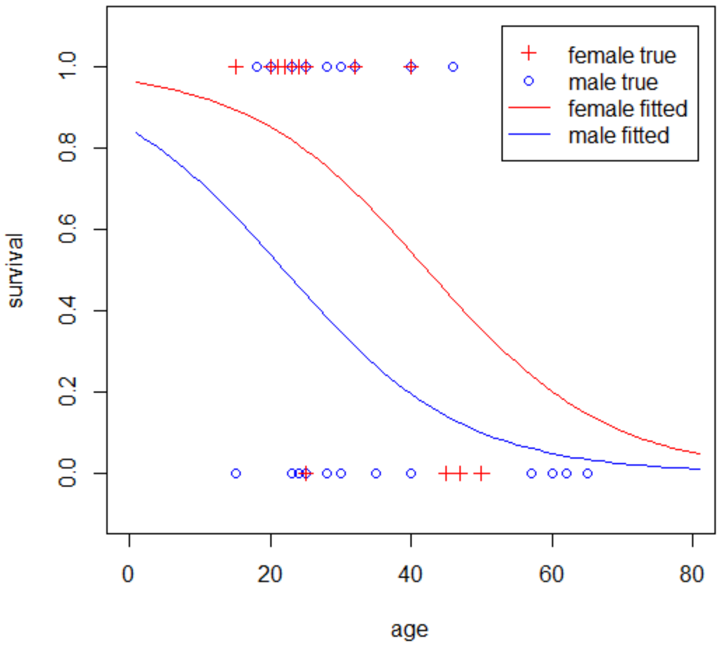

Figure 2: Logistic regression fitting

What implications may we have from the logistic regression model in (1)? Just like ordinary regression, the significance of the variables is critical. According to the regression report, we know that both coefficients are significantly nonzero. Their signs then provide useful implications. In particular, $-0.078\,age$ suggests that younger people will survive more likely, and

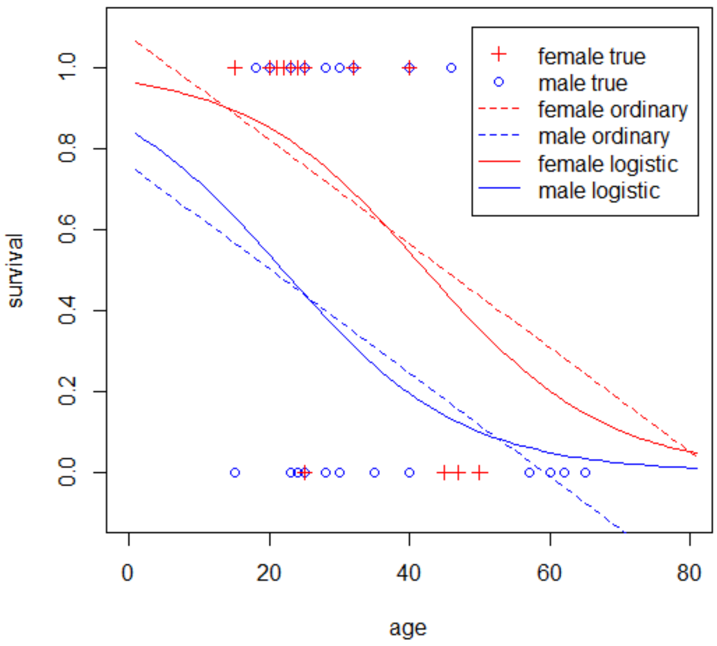

Figure 3: Comparison of the two types of regression models

$1.597\mathit{female}$ means that women will survive with a higher probability. Again, they fit our intuition, but with a better model, we are now more confident about our findings and conclusions.

In general, we read a logistic regression model in the following way. First, the *p-values* let us determine the significance of variables and select a good set of independent variables. The *signs* of significant coefficients then allow us to draw qualitative implications. Finally, we may use the *formula* itself to do prediction.

5 Model selection

Recall that in ordinary regression, we use R^2 and adjusted R^2 to assess the usefulness of a model. In logistic regression, unfortunately, we do not have R^2 and adjusted R^2. We then rely on a new concept *deviance* to evaluate a logistic regression model. For a given model, two types of deviances should be mentioned:

- The *null deviance* can be considered as the total estimation errors without using any independent variable. The null deviance is the same for all models (as long as the dependent variable is the same).

- The *residual deviance* can be considered as the total estimation errors by using the selected independent variables. Just like adding variables always increases R^2, adding additional variables into an existing model will always reduce the null deviance.

Ideally, the residual deviance should be small.[4]

Both the null and residual deviances are provided in the regression report. If we execute

[4] To be more rigorous, the residual deviance should also be close to its degree of freedom. This is beyond the scope of this article.

```
fitRight <- glm(d$survival ~ d$age + d$female, binomial)
summary(fitRight)
```

we will obtain

```
    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
```

This tells us that the null deviance is 61.827 and the residual deviance is 51.256. By trying some other combinations of independent variables, we may obtain their null and residual deviances. See Table 2 for a comparison. According to Table 2, one conclusion can be made: Because the residual deviance is smaller, using *age* only is better than using *female* only. We may make this conclusion because their numbers of variables are the same. Nevertheless, we cannot conclude that model 3 is better than model 1 simply by comparing their residual deviances. Their numbers of variables are different.

Model	Independent variable(s)	Null deviance	Residual deviance
1	*age*	61.827	56.291
2	*female*	61.827	57.286
3	*age*, age^2	61.827	55.822
4	*age*, *female*	61.827	51.256
5	*age*, *female*, *age* × *female*	61.827	47.346

Table 2: Null and residual deviances of various models

It remains to compare models with different numbers of variables. To take the number of variables into consideration, we may use *Akaike Information Criterion* (AIC). For each model, there is an associated AIC, which is also available in the regression report. Table 3 list the AICs of the five models.

Model	Independent variable(s)	Null deviance	Residual deviance	AIC
1	*age*	61.827	56.291	60.291
2	*female*	61.827	57.286	61.291
3	*age*, age^2	61.827	55.822	61.822
4	*age*, *female*	61.827	51.256	57.256
5	*age*, *female*, *age* × *female*	61.827	47.346	55.346

Table 3: AICs of various model

One important fact to note is that AIC can only be used to compare *nested* models, where two models are nested if one's variables are form a subset of the other's. With this in mind, we can now conclude that model 5 is better than model 4, model 4 is better than either model 1 or model 2, and model 3 is better than model 1. However, we cannot say that model 4 is better than model 3 because they are not nested.

In summary:[5]

- If two models have the same number of variables, compare their residual deviances.
- If two models are nested, compare their AICs.

[5] Other cases are beyond the scope of this article.