Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 14: Logistic Regression

Thursday 09:10-12:00, 資訊 140301

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

11.1 Regression Models with.Binary Response Variable*

11.2 Sigmoidal Response Functions for Binary Responses*

11.3 Simple Logistic Regression

1. If X is a random variable with _____, then

   _____

   and the probability mass function of this distribution

   _____.

2. The logit is the logarithm of the _____, where p = probability of a positive outcome (e.g., survived Titanic sinking)

   _____.

3. A formal statement of the _____: recall that when the response variable is _____, taking on the values _____ with probabilities _____ and _____, respectively, Y is a Bernoulli random variable with parameter _____.

4. We could state the simple logistic regression model in the usual form:

   _____

5. Since the distribution of the error term ε_i depends on the _____ distribution of the response Y_i, it is preferable to state the simple logistic regression model as: Y_i are independent Bernoulli random variables with expected values:

   _____ . (14.20)

6. The X observations are assumed to be known _____. Alternatively, if the X observations are random, $E\{Y_i\}$ is viewed as a _____, given the value of X_i.

Likelihood Function

1. Since each Y_i observation is an ordinary Bernoulli random variable, where:

   $$P(Y_j = 1) = \pi_i; \quad P(Y_j = 0) = 1 - \pi_i; \quad i = 1, \cdots, n.$$

   we can represent its probability distribution as follows:

   _____, $Y_i = 0, 1; \quad i = 1, \cdots, n.$ (14.21)

   Note that _____ and _____. Hence, $f_i(Y_i)$ simply represents the _____ that $Y_i = 1$ or 0.

2. Since the Y_i observations are independent, their joint probability function is:

   _____ . (14.22)

3. Find the maximum likelihood estimates by working with the logarithm of the joint probability function:

   $$\begin{aligned} \ln g(Y_1, \cdots, Y_n) &= \ln \prod_{i=1}^{n} f_i(Y_i) \\ &= \\ &\quad \_\_\_\_\_ \\ &= \\ &\quad \_\_\_\_\_ . \end{aligned}$$

4. Since $E\{Y_i\} = \pi_i$; for a binary variable, it follows from (14.20) that:

$$1 - \pi_i = \underline{\qquad\qquad\qquad\qquad} \tag{14.24}$$

5. Furthermore, from (14.l8a), we obtain:

(14.25)

$\underline{\qquad\qquad\qquad\qquad}$

6. Hence, log likelihood (14.23) can be expressed as follows:

$$\ln L(\beta_0, \beta_1) = \tag{14.26}$$

where $L(\beta_0, \beta_1)$ replaces $g(Y_1, \cdots, Y_n)$ to show explicitly that we now view this function as the likelihood function of the parameters to be estimated, given the sample observations.

Maximum Likelihood Estimation

1. The maximum likelihood estimates of β_0 and β_1 in the simple logistic regression model are those values of β_0 and β_1 that $\underline{\qquad\qquad}$ the log-likelihood function in (14.26).

2. $\underline{\qquad\qquad\qquad\qquad}$ exists for the values of β_0 and β_1, in 04.26) that maximize the log-likelihood function. Computer-intensive numerical search procedures are therefore required to find the maximum likelihood estimates b_0 and b_1.

3. Once the maximum likelihood estimates b_0 and b_1 are found, we substitute these values into the response function in (14.20) to obtain the fitted response function. We shall use π_i to denote the fitted value for the ith case:

.

$\underline{\qquad\qquad\qquad\qquad}$

4. The fitted logistic response function is as follows:

$\underline{\qquad\qquad\qquad\qquad}$

5. If we utilize the logit transformation in (14.18), we can express the fitted response function in (14.28) as follows:

$\underline{\qquad\qquad}$, $\underline{\qquad\qquad}$ (14.29)

We call (14.29) the $\underline{\qquad\qquad\qquad\qquad}$.

6. Once the fitted logistic response function has been obtained, the usual next steps are to _____ of the fitted response function and, if the fit is good, to make a variety of _____.

7. We shall postpone a discussion of how to examine the goodness of fit of a logistic response function and how to make inferences and predictions until we have considered the multiple logistic regression model with a number of predictor variables.

Example

1. A systems analyst studied the effect of computer programming experience on ability to complete within a specified time a complex programming task, including debugging. Twenty-five persons were selected for the study. They had varying amounts of programming experience (measured in months of experience), as shown in Table 14.1a column 1.

TABLE 14.1 Data and Maximum Likelihood Estimates—Programming Task Example.

(a) Data

Person i	(1) Months of Experience X_i	(2) Task Success Y_i	(3) Fitted Value $\hat{\pi}_i$	(4) Deviance Residual dev_i
1	14	0	.310	−.862
2	29	0	.835	−1.899
3	6	0	.110	−.483
...	...	...	...	...
23	28	1	.812	.646
24	22	1	.621	.976
25	8	1	.146	1.962

(b) Maximum Likelihood Estimates

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	Estimated Odds Ratio
β_0	−3.0597	1.259	—
β_1	.1615	.0650	1.175

2. All persons were given the same programming task, and the results of their success in the task are shown in column 2. The results are coded in binary fashion: $Y = 1$ if the task was completed successfully in the allotted time, and $Y = 0$ if the task was not complete d successfully.

3. (Figure 14.5) contains a scatter plot of the data. This plot is not too informative because of the nature of the response variable, other than to indicate that ability to complete the task successfully appears to increase with amount of experience. A lowess nonparametric response curve was fitted to the data and is also shown in Figure 14.5.

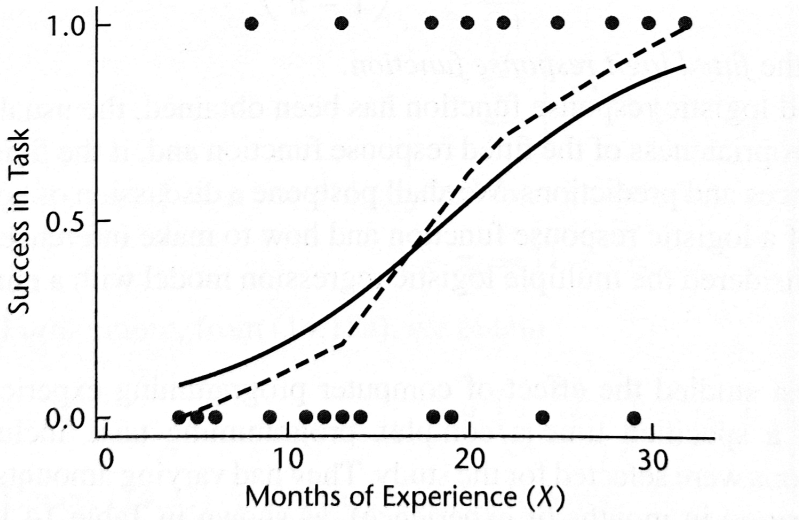

FIGURE 14.5 **Scatter Plot, Lowess Curve (dashed line), and Estimated Logistic Mean Response Function (solid line)—Programming Task Example.**

4. A _____ response function is clearly suggested by the _____ _____ fit. It was therefore decided to fit the _____ regression model (14.20).

5. A standard logistic regression package was run on the data. The results are contained in Table 14.1b. Since _____ and _____, the estimated logistic regression function:

_____.

6. This fitted value is the estimated probability that a person with 14 months experience ($X_1 = 14$) will successfully complete the programming task.

7. In addition to the lowess fit, Figure 14.5 also contains a plot of the fitted logistic response function, _____.

Interpretation of b_1

1. The interpretation of the estimated regression coefficient b_1 in the fitted logistic response function (14.30) is _____ of the slope in a linear regression model.

2. The reason is that the effect of a unit increase in X varies for the logistic regression model according to the _____ on the X scale.

3. An interpretation of b_1 is found in the property of the fitted logistic function that the estimated odds _____ are multiplied by _____ for any unit increase in X.

   (a) Consider the value of the fitted logit response function (14.29) at $X = X_j$:

   _____.

   The notation $\hat{\pi}'(X_j)$ indicates specifically the X level associated with the fitted value.

   (b) We also consider the value of the fitted logit response function at _____: The difference between the two fitted values is simply:

   _____.

   (c) Now according to (14.29a), $\hat{\pi}'(X_j)$ is the logarithm of the estimated odds when $X = X_j$; we shall denote it by $\log_e(\text{odds}_1)$. Similarly, $\hat{\pi}'(X_j+1)$ is the logarithm of the estimated odds when $X = X_j + 1$; we shall denote it by $\log_e(\text{odds}_2)$.

   _____.

   (d) Hence, the difference between the two fitted logit response values can be expressed as follows:

   $$\log_e(\text{odds}_2) - \log_e(\text{odds}_1) = \text{_____}$$

   (e) Taking _____ of each side, we see that the estimated ratio of the odds, called the _____ and denoted by $\widehat{OR}$, equals _____:

   _____ (14.31)

4. Example The programming task example.

   (a) We see from Figure 14.5 that the probability of success _____ with experience.

   (b) Specifically, Table 14.1b shows that the odds ratio is

   $$\widehat{OR} = \exp(b_1) = \exp(0.1615) = 1.175,$$

   so that the _____ by 17.5 percent with each additional month of experience.

   (c) Since a unit increase of one month is quite small, the estimated odds ratio of 1.175 may not adequately show the change in odds for a longer difference in time. In general, the estimated odds ratio when there is a _____ of X is _____.

   (d) For example, should we wish to compare individuals with relatively little experience to those with extensive experience, say 10 months versus 25 months so that $c = 15$, then the odds ratio would be estimated to be $\exp[15(0.1615)] = 11.3$. This indicates that the odds of completing the task increase over _____ for experienced persons compared to relatively inexperienced persons.

Supplementary

1. The 6 Assumptions of Logistic Regression

   (a) The response variable is _____.

   (b) The observations are _____.

   (c) There is _____ among explanatory variables.

   (d) There are _____.

   (e) There is a _____ between explanatory variables and the _____ Variable.

   (f) The sample size is sufficiently _____.

2. Assumptions of Logistic Regression vs. Linear Regression: In contrast to linear regression, logistic regression does not require:

(a) A linear relationship between the explanatory variable(s) and the response variable.

(b) The residuals of the model to be _____ distributed.

(c) The residuals to have _____, also known as _____.

😍 TA Class

- **Problems**: 14.7
- **Exercises**: none
- **Projects**: none