For large sample sizes, there is little difference between the t distribution and the standard normal distribution.

3. Approximate joint confidence intervals for several logistic regression parameters can be developed by the Bonferroni procedure. If g parameters are to be estimated with family confidence coefficient of approximately $1 - \alpha$, the joint Bonferroni confidence limits are:

$$b_k \pm Bs\{b_k\} \tag{14.56}$$

where:

$$B = z(1 - \alpha/2g) \tag{14.56a}$$

4. For power and sample size considerations in logistic regression modeling, see Reference 14.4 ■

Test whether Several $\beta_k = 0$: Likelihood Ratio Test

Frequently there is interest in determining whether a subset of the X variables in a multiple logistic regression model can be dropped, that is, in testing whether the associated regression coefficients β_k equal zero. The test procedure we shall employ is a general one for use with maximum likelihood estimation, and is analogous to the general linear test procedure for linear models. The test is called the *likelihood ratio test*, and, like the general linear test, is based on a comparison of full and reduced models. The test is valid for large sample sizes.

We begin with the full logistic model with response function:

$$\pi = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta}_F)]^{-1} \qquad \text{Full model} \tag{14.57}$$

where:

$$\mathbf{X}'\boldsymbol{\beta}_F = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

We then find the maximum likelihood estimates for the full model, now denoted by $\mathbf{b}_F$, and evaluate the likelihood function $L(\boldsymbol{\beta})$ when $\boldsymbol{\beta}_F = \mathbf{b}_F$. We shall denote this value of the likelihood function for the full model by $L(F)$.

The hypothesis we wish to test is:

$$\begin{aligned} &H_0\colon \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\ &H_a\colon \text{not all of the } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned} \tag{14.58}$$

where, for convenience, we arrange the model so that the last $p - q$ coefficients are those tested. The reduced logistic model therefore has the response function:

$$\pi = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta}_R)]^{-1} \qquad \text{Reduced model} \tag{14.59}$$

where:

$$\mathbf{X}'\boldsymbol{\beta}_R = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1}$$

Now we obtain the maximum likelihood estimates $\mathbf{b}_R$ for the reduced model and evaluate the likelihood function for the reduced model containing q parameters when $\boldsymbol{\beta}_R = \mathbf{b}_R$. We shall denote this value of the likelihood function for the reduced model by $L(R)$. It can be shown that $L(R)$ cannot exceed $L(F)$ since one cannot obtain a larger maximum for the likelihood function using a subset of the parameters.

The actual test statistic for the likelihood ratio test, denoted by G^2, is:

$$G^2 = -2\log_e\left[\frac{L(R)}{L(F)}\right] = -2[\log_e L(R) - \log_e L(F)] \qquad \textbf{(14.60)}$$

Note that if the ratio $L(R)/L(F)$ is small, indicating H_a is the appropriate conclusion, then G^2 is large. Thus, large values of G^2 lead to conclusion H_a.

Large-sample theory states that when n is large, G^2 is distributed approximately as $\chi^2(p-q)$ when H_0 in (14.58) holds. The degrees of freedom correspond to $df_R - df_F = (n-q)-(n-p) = p-q$. The appropriate decision rule therefore is:

$$\begin{aligned}&\text{If } G^2 \le \chi^2(1-\alpha; p-q), \text{ conclude } H_0\\ &\text{If } G^2 > \chi^2(1-\alpha; p-q), \text{ conclude } H_a\end{aligned} \qquad \textbf{(14.61)}$$

Example

In the disease outbreak example, the model building began with the three predictor variables that were considered *a priori* to be key explanatory variables—age, socioeconomic status, and city sector. A logistic regression model was fitted containing these three predictor variables and the log-likelihood for this model was obtained. Then tests were conducted to see whether a variable could be dropped from the model. First, age (X_1) was dropped from the logistic model and the log-likelihood for this reduced model was obtained. The results were:

$$L(F) = L(b_0, b_1, b_2, b_3, b_4) = -50.527 \qquad L(R) = L(b_0, b_2, b_3, b_4) = -53.102$$

Hence the required test statistic is:

$$G^2 = -2[\log_e L(R) - \log_e L(F)] = -2[-53.102 - (-50.527)] = 5.150$$

For $\alpha = .05$, we require $\chi^2(.95; 1) = 3.84$. Hence to test H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$, the appropriate decision rule is:

$$\begin{aligned}&\text{If } G^2 \le 3.84, \text{ conclude } H_0\\ &\text{If } G^2 > 3.84, \text{ conclude } H_a\end{aligned}$$

Since $G^2 = 5.15 \geq 3.84$, we conclude H_a, that X_1 should not be dropped from the model. The P-value of this test is .023.

Similar tests for socioeconomic status (X_2, X_3) and city sector (X_4) led to P-values of .55 and .001. The P-value for socioeconomic status suggests that it can be dropped from the model containing the other two predictor variables. However, since this variable was considered *a priori* to be important, additional analyses were conducted. When socioeconomic status is the only predictor in the logistic regression model, the P-value for the test whether this predictor variable is helpful is .16, suggesting marginal importance for this variable. In addition, the estimated regression coefficients for age and city sector and their estimated standard deviations are not appreciably affected by whether or not socioeconomic status is in the regression model. Hence, it was decided to keep socioeconomic status in the logistic regression model in view of its *a priori* importance.

The next question of concern was whether any two-factor interaction terms are required in the model. The full model now includes all possible two-factor interactions, in addition to the main effects, so that $\mathbf{X}'\boldsymbol{\beta}_F$ for this model is as follows:

$$\begin{aligned}\mathbf{X}'\boldsymbol{\beta}_F = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_1X_2 + \beta_6X_1X_3 \\ + \beta_7X_1X_4 + \beta_8X_2X_4 + \beta_9X_3X_4 \qquad \text{Full model}\end{aligned}$$

We wish to test:

$$\begin{aligned}&H_0\text{: } \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0 \\ &H_a\text{: not all } \beta_k \text{ in } H_0 \text{ equal zero}\end{aligned}$$

so that $\mathbf{X}'\boldsymbol{\beta}_R$ for the reduced model is:

$$\mathbf{X}'\boldsymbol{\beta}_R = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 \qquad \text{Reduced model}$$

A computer run of a multiple logistic regression package yielded:

$$\begin{aligned}L(F) &= -46.998 \\ L(R) &= -50.527 \\ G^2 &= -2[\log_e(R) - \log_e(F)] = 7.058\end{aligned}$$

If H_0 holds, G^2 follows approximately the chi-square distribution with 5 degrees of freedom. For $\alpha = .05$, we require $\chi^2(.95; 5) = 11.07$. Since $G^2 = 7.058 < 11.07$, we conclude H_0, that the two-factor interactions are not needed in the logistic regression model. The P-value of this test is .22. We note again that a logistic regression model without interaction terms is desirable, because otherwise $\exp(\beta_k)$ no longer can be interpreted as the odds ratio.

Thus, the fitted logistic regression model (14.46) was accepted as the model to be checked diagnostically and, finally, to be validated.

Comment

The Wald test for a single regression parameter in (14.53) is more versatile than the likelihood ratio test in (14.60). The latter can only be used to test H_0: $\beta_k = 0$, whereas the former can be used also for one-sided tests and for testing whether β_k equals some specified value other than zero. When testing H_0: $\beta_k = 0$, the two tests are not identical and may occasionally lead to different conclusions. For example, the Wald test P-value for dropping age when socioeconomic status and sector are in the model for the disease data set example is .0275; the P-value for the likelihood ratio test is .023. ■

14.6 Automatic Model Selection Methods

Several automatic model selection methods are available for building logistic regression models. These include all-possible-regressions and stepwise procedures. We begin with a discussion of criteria for model selection.

Model Selection Criteria

In the context of multiple linear regression models, we discussed the use of the followin model selection criteria in Chapter 9: R^2_p, $R^2_{a,p}$, C_p, AIC_p, SBC_p, and $PRESS_p$. For logistic regression modeling, the AIC_p and SBC_p criteria are easily adapted and are generall available in commercial software. For these reasons we will focus on the use of these t