UNIT 11 STATISTICAL INFERENCE IN LOGISTIC REGRESSION

Structure

11.1 INTRODUCTION

In the previous unit, we have discussed logistic regression model for determining the relationship of dichotomous response variable with one regressor variable. We have also explained how to fit a logistic regression model using maximum likelihood method of estimation applying iteratively reweighted least squares approach.

This unit deals with the inferential aspects of logistic model to draw inference about the population characteristics on the basis of information available from the sample. We define properties of the estimated regression coefficient in Sec. 11.2 whereas, we discuss the significance of the overall fitted logistic regression model using likelihood ratio test in Sec. 11.3. We consider testing the significance as well as computation of $(1-\alpha)100\%$ confidence intervals of the individual coefficients of the fitted logistic model in Sec. 11.4. We discuss the pseudo R^2 in Sec. 11.5 for the fitted logistic model.

In the next unit, you will learn about the fitting of multiple logistic regression models and its related statistical inference.

Objectives

After studying this unit, you should be able to:

- describe the properties of maximum likelihood estimators of parameters in logistic model;
- test the significance of the overall fitted logistic regression model;
- check the significance of the individual parameters of logistic regression;
- determine confidence interval of the parameters of logistic regression; and
- compute the pseudo R^2 for the fitted logistic models.

11.2 PROPERTIES OF ESTIMATED REGRESSION COEFFICIENTS

Let us assume that π_i be the final estimated value of the parameter β_0 and β_1 of logistic model. If the model assumptions are correct, we can show asymptotically:

1. The maximum likelihood estimators $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are unbiased estimators of β_0 and β_1 , respectively, i.e.,

$$E(\hat{\beta}_0^*) = \beta_0 \text{ and } E(\hat{\beta}_1^*) = \beta_1 \quad \text{... (1)}$$

2. We define the variances of $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ as:

$$\text{Var}(\hat{\beta}_0^*) = \frac{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)}{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)^2} \quad \text{... (2)}$$

$$\text{Var}(\hat{\beta}_1^*) = \frac{\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right)}{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)^2} \quad \text{... (3)}$$

The covariance of $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ is given as:

$$\text{Cov}(\hat{\beta}_0^*, \hat{\beta}_1^*) = \frac{-\left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)}{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)^2} \quad \text{.. (4)}$$

You may like to solve the following example to learn this concept.

Example 1: For the logistic model fitted in Example 5 of Unit 10, determine the standard errors of $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$. Also compute the covariance between them.

Solution: We use the values computed in Step 7 of Iteration 5 from the solution of Example 5 of Unit 10 and compute the variances of $\hat{\beta}_0^*$ and β_1^* using equations (2) and (3), respectively as:

$$\text{Var}(\hat{\beta}_0^*) = \frac{(34176.918167)}{(34176.918167)(22.396294) - (864.853059)^2} = 1.956826$$

$$\text{Var}(\hat{\beta}_1^*) = \frac{(22.396294)}{(34176.918167)(22.396294) - (864.853059)^2} = 0.001282$$

Thus, the standard errors of $\hat{\beta}_0^*$ and β_1^* are obtained as:

$SE(\hat{\beta}_0^*) = \sqrt{Var(\hat{\beta}_0^*)} = \sqrt{1.956826} = 1.398866$ and

$SE(\hat{\beta}_1^*) = \sqrt{Var(\hat{\beta}_1^*)} = \sqrt{0.001282} = 0.035809$

The covariance of $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ is calculated using equation (4) as:

$$Cov(\hat{\beta}_0^*, \hat{\beta}_1^*) = \frac{-864.853059}{(34176.918167)(22.396294) - (864.853059)^2} = -0.049518$$

You should solve the following exercises for practice.

E1) For the exercise given in **E6**, determine the variance and standard errors for the estimates of β_0 and β_1.

E2) For the exercise given in **E7**, obtain the variance and standard error of the estimated parameters of the fitted logistic model. Also compute the covariance between them.

11.3 GOODNESS OF FIT OF THE FITTED LOGISTIC MODEL

So far you have learnt how to fit logistic model by determining maximum likelihood estimator of β_0 and β_1 using iterative reweighted least squares method. After fitting the model, our next step will be assessing the significance of the fitted logistic model. Therefore, we should check the fitted logistic model using an appropriate measure of goodness-of-fit to test whether the deviation from fit is within some accepted limits or not. In this section, the goodness-of-fit measures based on the likelihood ratio and Hosmer-Lameshow tests are discussed.

As you are aware that the goodness-of-fit of a model describe by how much the fitted model deviate from the saturated model. The ideal model is usually referred to as the saturated model. For grouped data ($n_i > 1$), the maximum log-likelihood for saturated model (ignoring first term) using equation (40) can be defined as:

We are considering natural logarithm, i.e., $\log = \log_e$

$$(\log L)_S = \sum_{i=1}^{n} \{y_i \log \pi_i + (n_i - y_i) \log (1 - \pi_i)\} \quad \ldots (5)$$

In case of group data as discussed in Sec. 10.5, the maximum log-likelihood for the fitted full model after ignoring the first term is given as:

$$(\log L)_F = \sum_{i=1}^{n} \{y_i \log \hat{\pi}_i + (n_i - y_i) \log (1 - \hat{\pi}_i)\} \quad \ldots (6)$$

where $\hat{\pi}_i$ is the estimated value of π_i corresponding to the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$. We define the deviance statistic as:

$$D_F = -2 \log \left[\frac{(L)_F}{(L)_S} \right] \quad \ldots (7)$$

The quantity $\frac{(L)_F}{(L)_S}$ is known as likelihood ratio. The deviance statistic defined in equation (7) is used for testing the goodness-of-fit of the fitted model which is called likelihood ratio test.

In other words, we redefine the deviance 'D_F' given in equation (7) as two times of the difference between the maximum log-likelihood for saturated model and the fitted model. From equations (5) and (6), the deviance statistic is defined as:

$$D_F = -2[(\log L)_F - (\log L)_S] \quad \text{... (8)}$$

or, $$D_F = -2\sum_{i=1}^{n}\left[y_i \log_e\left(\frac{\hat{\pi}_i}{\pi_i}\right) + (n_i - y_i)\log_e\left(\frac{1-\hat{\pi}_i}{1-\pi_i}\right)\right] \quad \text{... (9)}$$

It is also known as model deviance for the fitted logistic model. Most of the computer software calculate the model deviance corresponding to each fitted logistic model. The null and alternative hypotheses for testing Deviance D_F can be formulated as:

H_0: The deviance is not significant, i.e., the logistic model fits good.

H_1: The deviance is significant, i.e., the logistic model does not fit good.

The degrees of freedom (df) is given by the difference between the number of parameters in two models. The number of parameters in the saturated model is n, while in the logistic model is 2. Hence the df for the model deviance is (n–2). To test the goodness-of-fit, the deviance is commonly compared with the Chi-square value with (n – 2) df. The smaller value of D_F shows the better fit of the model under the null hypothesis. The large value of deviance indicates that the fitted logistic model is not correct significantly. If the deviance will be significantly large, the model is considered to be a poor fit.

Decision rule: If $D_F \le \chi^2_{\alpha,n-2}$, we may not reject at the null hypothesis and the fitted model may be considered as adequate, otherwise we may reject H_0 and then we may look for other models which may give better fit.

It is to be noted that the value of log-likelihood for saturated model in case of ungrouped data (when the values of response variable are in the form of 0 or 1) will be zero as:

$$(\log_e L)_S = \sum_{i=1}^{n}\{y_i \log_e y_i + (1-y_i)\log_e(1-y_i)\} = 0 \quad \text{... (10)}$$

For evaluating the goodness-of-fit or significance of the regressor variable added in the model, we compare the deviances corresponding to fitted logistic model with or without adding the regressor variable in the model. The model without regressor variable (only intercept, β_0 model) can also be termed as null or reduced model. The maximum log-likelihood for null model (only intercept model) is obtained as:

$$(\log L)_N = \{N_1 \log(N_1) + N_0 \log(N_0) - N\log(N)\} \quad \text{… (11)}$$

where

= Total number of observations (1 and 0) of the response variable,

$N_1 = \sum_{i=1}^{n} y_i$ = Number of present outcome (1) in the response variable and

$N_0 = N - N_1$ = Number of absent outcome (0) in the response variable.

The deviance for the null model can be written as:

$$D_N = -2[(\log L)_R - (\log L)_S] \quad \ldots (12)$$

We define a statistic G based on Likelihood Ratio test which is the difference between the deviances of null and full models, as:

$$G = D_N - D_F \quad \ldots (13)$$

$$= -2[(\log L)_N - (\log L)_S] + 2[(\log L)_F - (\log L)_S]$$

$$= 2[(\log L)_F - (\log L)_N] \quad \ldots (14)$$

$$= 2\left[\begin{array}{l}\sum_{i=1}^{N}\{y_i \log \hat{\pi}_i + (n_i - y_i)\log(1-\hat{\pi}_i)\} \\ -\{N_1 \log(N_1) + N_0 \log(N_0) - N\log(N)\}\end{array}\right] \quad \ldots (15)$$

The null and alternative hypotheses for G statistic can be formulated as:

H_0: $\beta_1 = 0$, i.e, the fitted logistic model is not significant.

H_1: $\beta_1 \neq 0$, i.e, the fitted logistic model is significant.

Decision rule: The statistic G approximately follows Chi-square (χ^2) distribution with 1 degree of freedom. If $G \geq \chi^2_{1,\alpha}$, we may reject the null hypothesis and conclude that the regressor variable is contributing significantly to the model. Otherwise, we may not reject H_0 and the fitted model may be considered as an inadequate fit.

11.3.3 Hosmer-Lemeshow Test

Another method for assessing goodness-of-fit of the logistic regression model is the Hosmer-Lemeshow test. In the Hosmer-Lemeshow test, we divide the ungrouped data into g equal groups based on the ordered predicted probabilities (π_i's). The most commonly used method for creating groups is based on percentiles (or deciles in case of g=10). We first of all arrange the predicted probabilities in ascending order. The observations having lowest 10% predicted probabilities constitute the first group. The second group consist of the observations with next 10% lowest predicted probabilities. In the same way, our 10^{th} group consist of the observations with 10% highest predicted probabilities. In this way, we split the observations given in the data into 10 groups. We compute the total number of the observed ((from the given probabilities)) and expected (from the predicted probabilities) number of present outcomes (Y=1) and number of absent outcomes (Y=0) within each group. For this, we determine average of the respective probabilities in the group, and then multiply by the number of observations in the group for both presence and absence of the outcomes. It is to be remembered that each group should have enough observations. We define the Hosmer-Lameshow test statistic as:

$$C_{HL} = \sum_{k=0}^{1}\sum_{i=1}^{g} \frac{(O_{ki} - E_{ki})^2}{E_{ki}} \quad \ldots (16)$$

or, $$C_{HL} = \sum_{i=1}^{g}\left[\frac{(O_{1i} - E_{1i})^2}{E_{1i}} + \frac{(O_{0i} - E_{oi})^2}{E_{0i}}\right] \quad \ldots (17)$$

Where O_{1i} is number of present observed outcomes in the i^{th} group.

O_{0i} is the number of absent observed in the i^{th} group

E_{1i} is the number of expected present outcomes in the i^{th} group

E_{0i} is the number of expected absent outcomes in the i^{th} group.

We can also rewrite Hosmer-Lemeshow statistic equation (16) as:

$$C_{HL} = \sum_{i=1}^{g}\frac{\left(y_i - n_i'\hat{\bar{\pi}}_i\right)^2}{n_i'\hat{\pi}_i\left(1-\hat{\bar{\pi}}_i\right)} \quad \ldots (18)$$

where y_i is the number of observed outcomes in the i^{th} group.

n_1' is the total number of observations in the i^{th} group.

$\hat{\bar{\pi}}_i$ is the average of the predicted probabilities $(\hat{\pi}_i's)$ in the i^{th} group $n_i'\hat{\bar{\pi}}_i = \hat{y}_i$ is the expected number of outcomes in the i^{th} group.

If the number of observations is less, we may adjust the number of groups or can combine the groups. The Hosmer-Lemeshow C_{HL} - statistic follows the chi-square distribution with (g – 2) degrees of freedom. p-value can be the right tail probability of the corresponding chi-square distribution. The smaller p-value indicate the poor fit. The logistic model appears to fit well because we have no significant difference between the predicted and the observed data (i.e. the p-value is above 0.05).

The Hosmer-Lameshow test is also available in most of the computer software packages. In some software we can compute HL statistic by specifying the different values of g. for creating the group as described in this section, we need large data. We are not considering large ungrouped data here, so we will not be discussing creation of groups numerically here in this block as it will be difficult to handle manually. For applying the Hosmer-Lameshow test, we follow the steps given below.

Steps for Hosmer-Lameshow Test:

1. We estimate $\hat{\pi}_i$ for all n observations given in the data
2. We arrange $\hat{\pi}_i$ in ascending order from largest to smallest values.
3. We next divide the ordered values of $\hat{\pi}_i$ into g groups using percentiles. we generally consider g = 10, i.e., deciles.
4. We obtain the observed and expected total number of observation in each group and arrange them in a table.
5. After that we compute the Hosmer-Lameshow statistic using the formula given in equation (16, 17 or 18).
6. At last we compare the calculated values of Hosmer-Lameshow statistic with the tabulated value of χ^2 with g – 2 degrees of freedom and interpret the result.

Note that (i) when we have small number of observations in data, it is not possible to divide the data in suitable number of groups. (ii) when we have given the grouped data, we start from Step 4.

Let us solve the following example to understand the testing the significance of the fitted logistic model numerically.

Example 2: Test the significance of the logistic model fitted in Example 5 of Unit 10.

Solution: In this example, we are discussing the testing the significance of the fitted model using statistic(s) D_F, G and C_{HL}.

(i) Using model deviance D_F

From Table 6 given in the solution of Example 5 of Unit 10, we have

$y_1 = 7$, $y_2 = 4$, $y_3 = 10$, $y_4 = 14$, $\pi_1 = 0.21875$, $\pi_2 = 0.25$, $\pi_3 = 0.4$ and $\pi_4 = 0.4375$

From the Step 7 of Iteration 5 given in the solution of Example 5 of Unit 10, we have the predicted $\hat{\pi}_i$ as:

$$\hat{\pi}_1 = 0.216751, \hat{\pi}_2 = 0.284478, \hat{\pi}_3 = 0.363544 \text{ and } \hat{\pi}_4 = 0.450741$$

The maximum log-likelihood of saturated model (equation (5)) is computed as:

$$(\log L)_S = \sum_{i=1}^{n} \{y_i \log \pi_i + (n_i - y_i) \log (1 - \pi_i)\}$$

$$= -16.810282 - 8.997362 - 16.825292 - 21.930055$$

$$= -64.562991$$

The maximum log-likelihood of the fitted model given in equation (6) is obtained as:

$$(\log L)_F = \sum_{i=1}^{n} \{y_i \log \hat{\pi}_i + (n_i - y_i) \log(1 - \hat{\pi}_i)\}$$

$$= -16.810658 - 9.045312 - 16.896151 - 21.941407$$

$$= -64.693528$$

We compute the model deviance statistic given in equation (8) as:

$$D_F = -2[(\log L)_F - (\log L)_S]$$

$$= -2[-64.693528 + 64.562991] = -2(-0.130537) = 0.261074$$

Since df = n – 2 = 4 – 2 = 2, the tabulated value $\chi^2_{2,0.05} = 5.99$.

As $D_F < 5.99$, we may not reject the null hypothesis at 5 % level of significance and infer that the logistic model gives a reasonable fit.

(ii) Using G statistic

The maximum log-likelihood for reduced model (only intercept model) is obtained using equation (11) as:

$$(\log L)_R = \{N_1 \log(N_1) + N_0 \log(N_0) - N \log(N)\}$$

$$= 35 \log(35) + 70 \log(70) - 105 \log(105)$$

$$= 124.437182 + 297.3946674 - 488.665837 = -66.833988$$

We calculate the test statistic G using equation (14) as:

$$G = 2[(\log L)_F - (\log L)_R]$$

$$= 2\,(-64.693528 + 66.833988) = 2\,(2.140460) = 4.280919$$

The tabulated value $\chi^2_{1,0.05} = 3.84$.

Since G > 3.84, we may reject the null hypothesis at 5 % level of significance and conclude that the fitted logistic model gives a reasonable fit. The result indicates that the addition of regressor variable age to the logistic model is significantly better than the constant only model.

(iii) Using Hosmer-Lemeshow statistic

Since the given data is already grouped, we compute the HL statistic for g = 4.

From Table 6 given in the solution of Example 5 of Unit 10, we have

$$n_1' = n_1 = 32,\ n_2' = n_2 = 16,\ n_3' = n_3 = 25 \text{ and } n_4' = n_4 = 32$$

We compute the Hosmer-Lemeshow statistic CH_L using equations (18) as:

$$C_{HL} = \frac{(7 - 32\times 0.216751)^2}{32\times 0.216751(1-0.216751)} + \frac{(4 - 16\times 0.284478)^2}{16\times 0.284478(1-0.284478)}$$

$$+ \frac{(10 - 25\times 0.363544)^2}{25\times 0.363544(1-0.363544)} + \frac{(14 - 32\times 0.450741)^2}{32\times 0.450741(1-0.450741)}$$

$$= 0.000753 + 0.093438 + 0.143603 + 0.022663 = 0.260457$$

The tabulated of Chi-square at d.f. = g – 2 = 4 – 2 = 2 is $\chi^2_{2,0.05} = 5.99$

Since $CH_L < 5.99$, we may not reject the null hypothesis and the fitted model is seen as to be significantly fitted at 5% level of significance the basis of given data.

You may like to solve the following exercises for practice, before studying the next section.

E3) In case of ungrouped data, show that $(\log_e L)_S = 0$ for the saturated model.

E4) Test the goodness-of-fit of the fitted model in **E6** of **Unit 10** at 5% level of significance.

E5) For the exercise given in **E7** of **Unit 10**, test the goodness-of-fit of the fitted logistic model with respect to intercept only model at 1% level of significance.

11.4 STATISTICAL INFERENCE OF INDIVIDUAL COEFFICIENTS OF THE LOGISTIC MODEL

In this section, we discuss testing the significance as well as obtaining the confidence interval of the individual model coefficients:

11.4.1 Testing the significance

The Wald test and Score test are used to test the significance of the individual coefficients of the logistic model. In this block, we discuss only Wald test. For testing the null hypothesis $H_0 : \beta_j = 0$ verses the alternative hypothesis $H_1 : \beta_j \neq 0$; (j = 0 and 1), we define the Wald z-statistic as follows:

$$W_z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad \text{... (19)}$$

where Wald z-statistic follows a standard normal distribution, i.e., $W_z \sim N(0, 1)$.

Some literatures also use Wald chi-square statistic instead of Wald z-statistic and define as:

$$W_c = W_z^2 = \frac{\hat{\beta}_j^2}{Var(\hat{\beta}_j)} \quad \text{... (20)}$$

It is simply the square of the Wald z-statistic defined in equation (19) which follows Chi-square distribution with 1 degree of freedom. The tabulated values of $Z_{\alpha/2}$ or $\chi^2_{1,\alpha}$ at α% level of significance has been tabulated for various level of significance in Table I and II, respectively, given at the end of this block.

Decision Rule: If $|W_z| \geq Z_{\alpha/2}$ (or, If $|W_c| \geq \chi^2_{1,\alpha}$), we reject the null hypothesis at $\alpha\%$ level of significance. Otherwise, we do not reject it. In other words, we can say that there are sufficient evidences for rejection of the null hypothesis $H_0 : \beta_j = 0$.

It is to be noted that we can also use p-value approach to make a decision. If p-value is less than the significance level α, we may reject the null hypothesis. Otherwise, we do not reject it at $\alpha\%$ level of significance.

11.4.2 Confidence interval

In Unit 6 of Block 2, we have explained how to determine the confidence interval for linear regression models with the desired confidence level. We expect that the (1–α)100% times confidence interval will include the true value of the parameter. In the same way, we can determine the (1–α)100% lower and upper confidence limits for β_0 and β_1 based on respective Wald tests.

Note that we can also test the significance of the individual parameters of logistic model with the help of confidence interval. If the (1−α)100% confidence interval contains the value of the respective regression coefficient under null hypothesis, we do not reject the null hypothesis. Otherwise, we may reject the null hypothesis at α% level of significance.

(1– α)100% C.I. for β_0 is $((\hat{\beta}_0)_L, (\hat{\beta}_0)_U)$.

where $(\hat{\beta}_0)_L = \hat{\beta}_0 - z_{\alpha/2} SE(\hat{\beta}_0)$ and $(\hat{\beta}_0)_U = \hat{\beta}_0 + z_{\alpha/2} SE(\hat{\beta}_0)$ ... (21)

(1– α)100% C.I. for β_1 is $((\hat{\beta}_1)_L, (\hat{\beta}_1)_U)$.

where $(\hat{\beta}_1)_L = \hat{\beta}_1 - z_{\alpha/2} SE(\hat{\beta}_1)$ and $(\hat{\beta}_1)_U = \hat{\beta}_1 + z_{\alpha/2} SE(\hat{\beta}_1)$ ... (22)

So far in this section, we have discussed the testing of significance and the confidence interval estimation of the coefficients of logistic model. Let us now take up an example to illustrate the method.

Example 3: For the logistic model fitted on SBP data given in Example 5 of Unit 10, answers the following:

(i) Test the significance of the individual model coefficients β_0 and β_1 at 5% level of significance.

(ii) Obtain the 95% confidence intervals for the parameters of logistic model β_0 and β_1.

Solution: From the solution of Example 5 of Unit 10, we have

$\hat{\beta}_0^* = -3.458748$ and $\hat{\beta}_1^* = 0.072468$

From the solution of Example 1 of this unit, we have

$SE(\hat{\beta}_0^*) = 1.398866$ and $SE(\hat{\beta}_1^*) = 0.035809$

(i) $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$

We now compute the value of Wald z-statistic using equation (19) as:

$$W_z = \frac{\hat{\beta}_0^*}{SE(\hat{\beta}_0^*)} = \frac{-3.458748}{1.398866} = -2.472537$$

$$|W_z| = 2.472537$$

We can also determine Wald χ^2-statistic (equation (20)) as:

$$W_c = W_z^2 = (-2.472537)^2 = 6.1134394$$

At 5% level of significance, the tabulated Z and χ^2 values are:

$$Z_{0.025} = 1.96 \text{ and } \chi^2_{1,\,0.05} = 3.84$$

Since $|W_z| = 2.472537 > 1.96$ (or $W_c = 6.1134394 > 3.84$), we may reject the null hypothesis at 5% level of significance. Hence, we may conclude that there is sufficient evidence against H_0 and we may consider the value of intercept to be not equal to 0.

(ii) $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

We determine the value of Wald-statistic using equation (19) as:

$$W_z = \frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{0.072468}{0.035809} = 2.023719$$

$$|W_z| = 2.023719$$

We can also determine Wald χ^2-statistic using equation (20) as:

$$W_c = W_z^2 = (2.023719)^2 = 4.0954386$$

Since $|W_z| = 2.023719 > 1.96$ (or $W_c = 4.0954386 > 3.84$), we may reject the null hypothesis at 5% level of significance and conclude that the value of β_1 is significant. It means that the regressor variable X is contributing significantly to the model.

(i) We obtain the lower and upper confidence limits of β_0 as:

$$(\hat{\beta}_0)_L = \hat{\beta}_0 - z_{\alpha/2}\, SE(\hat{\beta}_0) = -3.458748 - 1.96 \times 1.398866 = -6.200525$$

$$(\hat{\beta}_0)_U = \hat{\beta}_0 + z_{\alpha/2}\, SE(\hat{\beta}_0) = -3.458748 + 1.96 \times 1.398866 = 0.716971$$

(ii) The lower and upper confidence limits of $\hat{\beta}_1$ can be determined as:

$$(\hat{\beta}_1)_L = \hat{\beta}_1 - z_{\alpha/2}\,SE(\hat{\beta}_1) = 0.072468 - 1.96 \times 0.035809 = 0.002282$$

$$(\hat{\beta}_1)_U = \hat{\beta}_1 + z_{\alpha/2}\,SE(\hat{\beta}_1) = 0.072468 + 1.96 \times 0.035809 = 0.142655$$

Thus, the 95% confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$ are (– 6.200525, 0.716971) and (0.002282, 0.142655), respectively.

You may now like to solve the following exercises to check your understanding:

E6) For the exercise given in **E6** of **Unit 10**, test the significance of the parameters of the logistic model at 5% level of significance using Wald z-statistic and determine the 95% confidence limits of the β_0 and β_1.

E7) For the exercise given in **E7** of **Unit 10**, test the significance of β_1 at 1% level of significance using Wald χ^2-statistic and compute the 99% confidence limits of β_1.

11.5 PSEUDO R-SQUARED

You must be thinking of R^2 (coefficient of determination) computed in case of linear regression models which we have discussed in Block 2 of this course. Like R^2 used in case of linear regression models, we compute **pseudo R-squared** in case of logistic models. The various formulae are given in the literatures to determine the **pseudo R-squared**. In this course, we discuss three of them namely "McFadden", "Cox and Snell" and "Nagelkerke" **pseudo R-squared.**

The McFadden pseudo R^2 is defined as:

$$R^2_{MF} = 1 - \frac{(\log L)_F}{(\log L)_N} \quad \ldots (23)$$

Where $(\log L)_F$ and $(\log L)_N$ are the full and null log likelihood functions defined in equations (6) and (11), respectively.

We express the Cox and Snell pseudo R^2 in terms of log likelihood as:

$$R^2_{cs} = 1 - \exp\left(\frac{2}{n}\{(\log L)_N - (\log L)_F\}\right) \quad \ldots (24)$$

In terms of likelihood function, we define

$$R^2_{cs} = 1 - \left\{\frac{(L)_N}{(L)_F}\right\}^{2/n} \quad \ldots (25)$$

Where n = Total number of observations or sample size.

$(L)_F$ and $(L)_N$ are the likelihood function of full and null models

The Nagelkerke's pseudo R^2 can be determined as:

$$R^2_N = \frac{1 - \left\{\frac{(L)_N}{(L)_F}\right\}^{2/n}}{1 - \{(L)_N\}^{2/n}} \quad \ldots (26)$$

or, $$R_N^2 = \frac{R_{cs}^2}{1-\{(L)_N\}^{2/n}} \quad \ldots (27)$$

In terms of log-likelihood, R_N^2 is defined as:

$$R_N^2 = \frac{R_{cs}^2}{1-\exp\left(\frac{2}{n}(\log L)_N\right)} \quad \ldots (28)$$

You may note that Nagelkerke pseudo R-squared is the corrected version of the Cox and Snell pseudo R-squared.

Remember that the logic and method of calculation used in logistic regression is different than that used for linear regression models in ordinary least squared case. Note that the value of **pseudo R-squared** is much smaller than the value of R^2. We generally use **pseudo R-squared** to compare different fitted logistic models of the same type which are fitted on the same data. The model with higher pseudo R-squared than the other models suggesting a better fit.
It indicates that the model with higher pseudo R-squared better fits to the data and better predicts the probability of outcome. The various computer software provides different types of pseudo R-squared values.

Example 4: For the logistic model fitted on SBP data given in Example 5 of Unit 10, determine the McFadden, Cox and Snell and Nagelkerke **pseudo R-squared**.

Solution: From the solution of Example 2, we have

$$(\log L)_F = -64.693528 \text{ and } (\log L)_N = -66.833988$$

We compute the McFadden pseudo R^2 using equation (23) as:

$$R_{MF}^2 = 1-\frac{(\log L)_F}{(\log L)_N}$$

$$= 1-\frac{(-64.693528)}{(-66.833988)} = 1-0.967973 = 0.032027$$

From equation (24), Cox and Snell pseudo R^2 is determined as:

$$R_{cs}^2 = 1-\exp\left(\frac{2}{n}\{(\log L)_N - (\log L)_F\}\right)$$

$$= 1-e^{\frac{2}{105}\{-66.833988+64.693528\}}$$

$$= 1-e^{\frac{2}{105}\{-2.140460\}} = 1-e^{-0.040771}$$

$$= 1-0.960049 = 0.039951$$

Nagelkerke's pseudo R^2 can be obtained using equation (28) as:

$$R_N^2 = \frac{R_{cs}^2}{1-\exp\left(\frac{2}{n}(\log L)_N\right)} = \frac{0.039951}{1-e^{\frac{2\times(-66.833988)}{105}}}$$

$$= \frac{0.039951}{1-e^{-1.273028}} = \frac{0.039951}{1-0.279983} = \frac{0.039951}{0.720018} = 0.055486$$

Thus, the values of McFadden, Cox and Snell and Nagelkerke **pseudo R-squared** are 0.032027, 0.039951 and 0.055486, respectively, for the logistic model fitted in Example 5.

You may now like to solve the following exercises to check your understanding:

E8) For the exercise given in **E6** of **Unit 10**, obtain the values of McFadden, Cox and Snell and Nagelkerke **pseudo R-squared**

E9) For the exercise given in **E7** of **Unit 10**, compute the values of McFadden, Cox and Snell and Nagelkerke **pseudo R-squared**.

So far, you have learnt how to fit a simple logistic regression model and its related statistical inference. In the next unit, we shall explain the fitting of multiple logistic regression model.

We now end this unit by giving a summary of what you have learnt in it.

11.6 SUMMARY

In this unit we have discussed:

1. The variances of the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are given as:

$$\text{Var}(\hat{\beta}_0) = \frac{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)}{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)^2}$$

$$\text{and } \text{Var}(\hat{\beta}_1) = \frac{\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right)}{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)^2}$$

2. The covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$ is given as:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)}{\left(\sum_{i=1}^{n} \frac{x_i^2}{V_{ii}}\right)\left(\sum_{i=1}^{n} \frac{1}{V_{ii}}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{V_{ii}}\right)^2}$$

3. The deviance 'D_F' is defined as two times of the difference between the maximum log-likelihood for saturated model and the fitted model which is given as:

$$D_F = -2[(\log L)_F - (\log L)_S],$$

$$D_F = -2\sum_{i=1}^{n}\left[y_i \log_e\left(\frac{\hat{\pi}_i}{\pi_i}\right) + (n_i - y_i)\log_e\left(\frac{1-\hat{\pi}_i}{1-\pi_i}\right)\right]$$

The deviance D_F is commonly compared with the Chi-square value with (n – 2) df.

4. We define a Likelihood Ratio (LR) test statistic G as difference between deviances of the null and fitted models as:

$$G = D_R - D_F$$

$$G = 2\left[\sum_{i=1}^{N}\{y_i \log \hat{\pi}_i + (n_i - y_i)\log(1-\hat{\pi}_i)\} - \{N_1 \log(N_1) + N_0 \log(N_0) - N\log(N)\}\right]$$

The statistic G approximately follows Chi-square (χ^2) distribution with 1 degree of freedom.

5. The Hosmer-Lemeshow test statistic is given as:

$$C_{HL} = \sum_{i=1}^{g} \frac{\left(y_i - n_i' \hat{\bar{\pi}}_i\right)^2}{n_i' \hat{\pi}_i \left(1 - \hat{\bar{\pi}}_i\right)}$$

The C_{HL}-statistic follows the chi-square distribution with (g – 2) degrees of freedom.

6. The Wald z and χ^2-statistic(s) for testing the $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$; (j = 0 and 1) are, respectively, given as:

$$W_z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad \text{and} \quad W_c = W_z^2 = \frac{\hat{\beta}_j^2}{Var(\hat{\beta}_j)}$$

7. We define the lower and upper limits of $(1-\alpha)100\%$ confidence interval $((\hat{\beta}_j)_L, (\hat{\beta}_j)_U)$ of $\beta_j; (j = 0 \text{ and } 1)$ as:

$$(\hat{\beta}_j)_L = \hat{\beta}_j - z_{\alpha/2}\, SE(\hat{\beta}_j) \text{ and } (\hat{\beta}_j)_U = \hat{\beta}_j + z_{\alpha/2}\, SE(\hat{\beta}_j)$$

8. The McFadden's pseudo R^2 is defined as:

$$R^2_{MF} = 1 - \frac{(\log L)_F}{(\log L)_N}$$

9. We define the Cox and Snell's pseudo R^2 in terms of log likelihood as:

$$R^2_{cs} = 1 - \exp\left(\frac{2}{n}\{(\log L)_N - (\log L)_F\}\right)$$

10. The Nagelkerke's pseudo R^2 is determined as:

$$R^2_N = \frac{R^2_{cs}}{1 - \exp\left(\frac{2}{n}(\log L)_N\right)}$$

11.7 SOLUTION/ANSWERS

E1) We use the values computed Step 6 from the solution of **E6** of **Unit 10** and compute the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ using equations (2) and (3) as:

$$Var(\hat{\beta}_0) = 0.079785 \text{ and } Var(\hat{\beta}_1) = 0.000255$$

The standard errors of $\hat{\beta}_0$ and β_1 are:

$SE(\hat{\beta}_0) = \sqrt{0.079785} = 0.282462$ and $SE(\hat{\beta}_1) = \sqrt{0.000255} = 0.015967$

E2) We use the values computed in Table 4 from the solution of **E7** of **Unit 10** and compute the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ using equations (2) and (3) as:

$$Var(\hat{\beta}_0) = 8.412958 \text{ and } Var(\hat{\beta}_1) = 0.006025$$

The standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$SE(\hat{\beta}_0) = \sqrt{8.412958} = 2.900510 \text{ and } SE(\hat{\beta}_1) = \sqrt{0.006025} = 0.077619$$

From using equation (4), the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ is computed as:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -0.215972$$

E3) For the saturated model, the likelihood function is

$$(L)_S = \prod_{i=1}^{n} (\pi_i)^{y_i} (1-\pi_i)^{(1-y_i)}$$

In case of ungrouped data, we have the values of y_i is either 1 or 0. From the definition of saturated model, we have $\pi_i = \frac{y_i}{n_i} = y_i$. Thus

$$(L)_S = 1 \Rightarrow (\log_e L)_S = 0$$

E4) From the data given in **E6** of **Unit 10**, we have

$y_1 = 24$, $y_2 = 18$, $y_3 = 12$, $y_4 = 20$, $y_5 = 26$, $\pi_1 = 0.4$, $\pi_2 = 0.375$,

$\pi_3 = 0.3$, $\pi_4 = 0.25$ and $\pi_5 = 0.25$

From the Step 3 given in the solution of **E6** of **Unit 10**, we have

$\hat{\pi}_1 = 0.400091$, $\hat{\pi}_2 = 0.355118$, $\hat{\pi}_3 = 0.312566$,

$\hat{\pi}_4 = 0.272956$ and $\hat{\pi}_5 = 0.236637$

The maximum log-likelihood for saturated model is:

$$(\log_e L)_S = -40.380700 - 31.755035 - 24.434572 - 44.986812 - 58.482855$$
$$= -200.039974$$

The maximum log-likelihood for the fitted model is:

$$(\log L)_F = -40.380701 - 31.796131 - 24.449381 - 45.094987 - 58.533615$$
$$= -200.254815$$

We compute the model deviance statistic as:

$$D_F = -2[-200.254815 + 200.039974] = -2(-0.214841) = 0.429683$$

Since df = n – 2 = 5 – 2 = 3, the tabulated value $\chi^2_{3,0.05} = 7.81$.

As $D_F < 7.81$, we may not reject the null hypothesis at 5 % level of significance and infer that the logistic model gives an adequate fit.

The maximum log-likelihood for only intercept model is given as:

$$(\log L)_R = 100\log(100) + 232\log(232) - 332\log(332) = -203.144721$$

We calculate the test statistic G as:

$$G = 2\,(-200.254815 + 203.144721) = 5.779811$$

The tabulated value $\chi^2_{1,0.05} = 3.84$.

Since G > 3.84, we may reject the null hypothesis at 5 % level of significance and say that the model with regressor variable (amount of dose) is significantly contributing than the constant only model.

Since the given data is already grouped in 5 classes, we have

$$g = 5,\ n_1' = n_1 = 60,\ n_2' = n_2 = 48,\ n_3' = n_3 = 40,$$

$$n_4' = n_4 = 80 \text{ and } n_5' = n_5 = 104$$

We compute the Hosmer-Lemeshow statistic CH_L as:

$$C_{HL} = \frac{(24 - 60\times 0.400091)^2}{60\times 0.400091(1-0.400091)} + \frac{(18 - 48\times 0.355118)^2}{48\times 0.355118(1-0.355118)}$$

$$+ \frac{(12 - 40\times 0.312566)^2}{40\times 0.312566(1-0.312566)} + \frac{(20 - 80\times 0.272956)^2}{80\times 0.272956(1-0.272956)}$$

$$+ \frac{(26 - 104\times 0.236637)^2}{104\times 0.236637(1-0.236637)}$$

$$= 0.000002 + 0.082855 + 0.029398 + 0.212431 + 0.102813$$

$$= 0.427498$$

The tabulated of Chi-square at d.f. = g – 2 = 5 – 2 = 3 is $\chi^2_{2,0.05} = 7.81$

Since $CH_L < 7.81$, we may not reject the null hypothesis and the fitted model is seen as to be significantly fitted at 5% level of significance the basis of given data.

E5) From the data given in **E7** of **Unit 10**, we have

$$y_1 = 0,\ y_2 = 0,\ y_3 = 1,\ y_4 = 0,\ y_5 = 1,\ y_6 = 1,\ N_1 = 3,\ N_0 = 3 \text{ and } N = 6$$

From the solution of Table 4 of **E7** of **Unit 10**, we have

$$\hat{\pi}_1 = 0.484442,\ \hat{\pi}_2 = 0.558952,\ \hat{\pi}_3 = 0.540441,$$

$$\hat{\pi}_4 = 0.447238,\ \hat{\pi}_5 = 0.465792 \text{ and } \hat{\pi}_6 = 0.503134$$

The maximum log-likelihood for the fitted model is:

$$(\log L)_F = -0.662505 - 0.818602 - 0.615369$$

$$- 0.592828 - 0.764016 - 0.686898$$

$$= -4.140217$$

The maximum log-likelihood for only intercept model is given as:

$$(\log L)_N = 3\log(3) + 3\log(3) - 6\log(6) = -4.158883$$

We calculate the test statistic G as:

$$G = 2\,(-4.140217 + 4.158883) = 0.037331$$

The tabulated value $\chi^2_{1,0.05} = 6.63$.

Since G < 6.63, we may not reject the null hypothesis at 1 % level of significance and conclude that the model with regressor variable duration of walk is not significantly contributing than the constant only model on the basis of given data.

Note that the size of given data in very small (n = 6) which cannot be classified further in groups. So in this case, it is not feasible to apply Hosmer-Lemeshow goodness-of-fit test.

E6) From the solutions of **E6** of **Unit 10** and **E3**, we have

$\hat{\beta}_0^* = -0.213558$, $\hat{\beta}_1^* = -0.038306$, $SE(\hat{\beta}_0^*) = 0.282462$ and $SE(\hat{\beta}_1^*) = 0.015967$

(i) $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$

The value of Wald z-statistic is:

$$W_z = \frac{\hat{\beta}_0^*}{SE(\hat{\beta}_0^*)} = \frac{-0.213558}{0.282462} = -0.756060$$

Since $|W_z| < 1.96$, we may not reject the null hypothesis at 5% level of significance for β_0.

(ii) $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

The Wald χ^2-statistic is computed as:

$$W_z = \frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{-0.038306}{0.015967} = -2.399147$$

Since $|W_z| > 1.96$, we may reject the null hypothesis at 5% level of significance and conclude that the coefficient β_1 is significantly contributing to the model.

(iii) We obtain the lower and upper confidence limits of β_0 as:

$$(\hat{\beta}_0)_L = \hat{\beta}_0 - z_{\alpha/2} SE(\hat{\beta}_0) = -0.213558 - 1.96 \times 0.282462 = -0.767183$$

$$(\hat{\beta}_0)_U = \hat{\beta}_0 + z_{\alpha/2} SE(\hat{\beta}_0) = -0.213558 + 1.96 \times 0.282462 = 0.340067$$

(iv) The lower and upper confidence limits of $\hat{\beta}_1$ can be determined as:

$$(\hat{\beta}_1)_L = \hat{\beta}_1 - z_{\alpha/2} SE(\hat{\beta}_1) = -0.038306 - 1.96 \times 0.015967$$

$$= -2.399147$$

$$(\hat{\beta}_1)_U = \hat{\beta}_1 + z_{\alpha/2} SE(\hat{\beta}_1) = -0.038306 + 1.96 \times 0.015967$$

$$= -0.069601$$

E7) From the solutions of **E7** of **Unit 10** and **E3**, we have

$\hat{\beta}_0^* = 0.536077$ and $\hat{\beta}_1^* = -0.014958$

$\hat{\beta}_1^* = -0.014958$, and $Var(\hat{\beta}_1^*) = 0.006025$

(i) $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

We determine Wald χ^2-statistic as:

$$W_c = \frac{(\hat{\beta}_1^*)^2}{\text{Var}(\hat{\beta}_1^*)} = \frac{(-0.014958)^2}{0.006025} = 0.037138$$

The tabulated χ^2 value is $\chi^2_{1,\,0.01} = 6.63$

Since $W_c < 6.63$, we may not reject the null hypothesis at 1% level of significance and conclude that the walking duration is not contributing significantly to the model.

(v) The lower and upper confidence limits of $\hat{\beta}_1$ can be determined as:

$$(\hat{\beta}_1)_L = \hat{\beta}_1 - z_{\alpha/2}\,\text{SE}(\hat{\beta}_1) = -0.014958 - 2.58 \times 0.077619 = -0.215215$$

$$(\hat{\beta}_1)_U = \hat{\beta}_1 + z_{\alpha/2}\,\text{SE}(\hat{\beta}_1) = -0.014958 + 2.58 \times 0.077619 = 0.185298$$

E8) From the solution of **E4**, we have

$$(\log L)_F = -200.254815 \text{ and } (\log L)_N = -203.144721$$

We compute the McFadden pseudo R^2 as:

$$R^2_{MF} = 1 - \frac{(-200.254815)}{(-203.144721)} = 1 - 0.985774 = 0.014226$$

Cox and Snell pseudo R^2 is determined as:

$$R^2_{cs} = 1 - e^{\frac{2}{332}\{-203.144721 + 200.254815\}}$$

$$= 1 - e^{-0.017409} = 1 - 0.982742 = 0.017258$$

Nagelkerke's pseudo R^2 can be obtained as:

$$R^2_N = \frac{0.017258}{1 - e^{\frac{2\times(-203.144721)}{332}}} = \frac{0.017258}{1 - e^{-1.22376}} = \frac{0.017258}{1 - 0.294121} = \frac{0.017258}{0.705879}$$

$$= 0.024450$$

Thus, the values of McFadden, Cox and Snell and Nagelkerke **pseudo R-squared** are 0.014226, 0.017258 and 0.024450, respectively.

E9) From the solution of **E5**, we have

$$(\log L)_F = -4.140217 \text{ and } (\log L)_N = -4.158883$$

We compute the McFadden pseudo R^2 as:

$$R^2_{MF} = 1 - \frac{(-4.140217)}{(-4.158883)} = 1 - 0.995512 = 0.004488$$

Cox and Snell pseudo R^2 is determined as:

$$R^2_{cs} = 1 - e^{\frac{2}{6}\{-4.158883 + 4.140217\}}$$

$$= 1 - e^{-0.006222} = 1 - 0.993797 = 0.006203$$

Nagelkerke's pseudo R^2 can be obtained as:

$$R^2_N = \frac{0.006203}{1 - e^{\frac{2\times(-4.158883)}{6}}} = \frac{0.006203}{1 - e^{-1.38629}} = \frac{0.006203}{1 - 0.25} = \frac{0.006203}{0.75} = 0.008270$$

The values of McFadden, Cox and Snell and Nagelkerke **pseudo R-squared** are 0.004488, 0.006203 and 0.008270, respectively.