The general form for the probability density function (pdf) of a logistic random variable, Lg(mean μ, variance σ^2), X is given by g(x; μ, σ) =

$$\frac{\pi}{\sigma\sqrt{3}}\ \frac{e^{-\pi(x-\mu)/\sigma\sqrt{3}}}{[1+e^{-\pi(x-\mu)/\sigma\sqrt{3}}]^2}\ ,\ -\infty < X < \infty\ \ ,\ -\infty < \mu < \infty\ ,\ \sigma_x > 0 \qquad (19)$$

Making the transformation Z = $\frac{\pi(X-\mu_x)}{\sigma_x\sqrt{3}}$ **in Eq. (19), the most common form of the logistic probability density function is obtained by letting G(x) and F(x) represent the cdf (cumulative distribution function) of X and Z, respectively. As a result [letting** $\mu = \mu_x$ & $\sigma = \sigma_x = \sqrt{V(X)}$ **], we obtain (x =** $\mu + z\sigma\sqrt{3}/\pi$ **)**

F(z) = Pr(Z ≤ z) = Pr($\frac{\pi(X-\mu)}{\sigma\sqrt{3}} \le z$ **) = Pr(** $X \le \mu + z\sigma\sqrt{3}/\pi$ **)=** $G_X(\mu + z\sigma\sqrt{3}/\pi)$

$$\rightarrow \text{f(z)} = \frac{dF(z)}{dz} = \frac{dG_X(\mu+z\sigma\sqrt{3}/\pi)}{dz} = \frac{dG_X(\mu+z\sigma\sqrt{3}/\pi)}{dx}\times\frac{dx}{dz} = \frac{dG_X(x)}{dx}\times\frac{dx}{dz}$$

$$= g(x)\times\frac{dx}{dz} = \frac{\pi}{\sigma\sqrt{3}}\ \frac{e^{-\pi(x-\mu)/\sigma\sqrt{3}}}{[1+e^{-\pi(x-\mu)/\sigma\sqrt{3}}]^2}\times\frac{\sigma\sqrt{3}}{\pi} = \frac{e^{-z}}{[1+e^{-z}]^2}$$

Thus, the standard (but not standardized) form the logistic pdf is given by

$$\text{f(z)} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{e^{z}}{(1+e^{z})^2} = \frac{1}{4}\text{Sech}^2(z/2),\quad -\infty < z < \infty \qquad (20)$$

where $\text{Cosh}(z) = (e^z + e^{-z})/2$, $\text{Sech}(z) = [(e^z + e^{-z})/2]^{-1} = 2/(e^z + e^{-z})$,

$\text{Sinh}(z) = (e^z - e^{-z})/2$, **and Tanh(z) =** $(e^z - e^{-z})/(e^z + e^{-z})$**. The expression**

f(z) = $\frac{1}{4}\text{sech}^2\frac{z}{2}$ **in Eq. (20) is obtained from the fact that** $\text{Sech}(z) = \frac{2}{e^z + e^{-z}}$ **=**

$$\frac{2e^z}{e^{2z}+1} = \frac{2e^z}{1+e^{2z}} \rightarrow \text{Sech}(z/2) = \frac{2e^{z/2}}{1+e^z} \rightarrow \frac{1}{2}\text{Sech}(z/2) = \frac{e^{z/2}}{1+e^z} \rightarrow$$

$\frac{1}{4}\text{Sech}^2(z/2) = [\frac{e^{z/2}}{1+e^z}]^2 \rightarrow \frac{1}{4}\text{Sech}^2(z/2) = \frac{e^z}{(1+e^z)^2}$ = f(z). Similarly, it can be verified that Cosh(z) = $2\cosh^2(z/2) - 1$, or Cosh(2z) = $2\cosh^2(z) - 1$, $\text{Cosh}^2(z) - \text{Sinh}^2(z) = 1$, $\text{Sech}^2(z) = 1 - \text{Tanh}^2(z)$ and Sinh(2z) = 2Sinh(z)Cosh(z).

Further, $\int_{-\infty}^{\infty} \frac{e^{-z}\,dz}{(1+e^{-z})^2} = \left[(1+e^{-z})^{-1}\right]_{-\infty}^{+\infty}$ = 1 shows that the integrand is a pdf over the range $-\infty < z < \infty$, and it can be verified from Eq. (20) that E(Z) = $\int_{-\infty}^{\infty} \frac{z\,e^{-z}\,dz}{(1+e^{-z})^2}$ = 0, V(Z) = $\frac{\pi^2}{3} = \mu_2$, the skewness $\alpha_3 = 0$ and the kurtosis $\beta_4 = \frac{\mu_4}{(\mu_2)^2} - 3 = \alpha_4 - 3$ = 1.20, where $\mu_4 = E(Z^4)$ and $\mu_r = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^r\right]$ represents the r^{th} central moment of a rv X. The pdf in (20) is symmetrical about zero, i.e., E(Z) = $z_{0.50}$ = The mode = 0. Thus, the graph of the standard logistic pdf resembles the standard normal density but with $\sigma_Z = \pi/\sqrt{3}$, thicker tails (normal kurtosis is zero) and less peaked in the middle because the height of the density at z = 0 for the standard logistic is 0.25 while it is equal to $1/\sqrt{2\pi}$ = 0.3989423 for the N(0, 1) pdf. Note that the standard form of the logistic does not have a variance of 1, but the mean is zero. The cdf of Z is [for both forms of the pdf in Eq. (20)] given by

$$F(z) = \int_{-\infty}^{z} f(u)du = \int_{-\infty}^{z} \frac{e^{-u}}{(1+e^{-u})^2}du = \left[(1+e^{-u})^{-1}\right]_{-\infty}^{z} = \frac{1}{1+e^{-z}}, -\infty < z < \infty \qquad (21)$$

Inverting Eq. (21) for z in terms of the cdf yields $1 + e^{-z}$ =1/F(z) $\rightarrow$ e^{-z} = – 1 + 1/F(z) $\rightarrow$ –z = ln[– 1 + 1/F(z)] $\rightarrow$ –z = ln[$\frac{1 - F(z)}{F(z)}$] $\rightarrow$ z = ln[$\frac{F(z)}{1 - F(z)}$] $\rightarrow$

$$z_p = \ln[\frac{p}{1-p}] \qquad (22a)$$

Eq. (22a) gives the percentile function of the standard form of the logistic density and again verifies that the 50th percentile (p = 0.50) is equal to $z_{0.50}$ = ln(0.5/0.5) = 0 and the 95th percentile, with standard deviation $\pi / \sqrt{3}$, is equal to $z_{0.95}$ = ln(0.95/0.05) = 2.94444.

Now, consider a dichotomous outcome variable with occurrence Pr of p when y = 1 and failure Pr equal to (1 –p) when y = 0, where p is a function of a continuous variable x such as height, weight, age, length of time at a job, temperature, etc., and the dichotomous variable may represent the presence (y =1) or absence (y = 0) of coronary disease, or injured on the job (y =1) or not injured on the job (y = 0), needs repair (y=1) or does not need repair (y=0). That is, in Logistic Regression, the response variable is dichotomous (0 or 1) and further we assume that p = Pr (Y =1) is a function of one (or more) continuous explanatory variable(s), i.e., in fact p = Pr (Y =1| x). For example, the probability p that a person has coronary disease is related to his/her age x, i.e., p actually is p(x), and in this case, p(x) is an increasing function of x. As the age of a person increases, the chance that the person has coronary heart disease increases. In Logistic Regression literature the occurrence Pr of the dichotomous outcome is generally referred to as $\pi(x)$ but for simplicity of notation I will use p(x) to denote this success Pr, i.e., p(x) represents the Pr that a Bernoulli rv (success/failure; presence/absence; reliable/unreliable, etc) occurs with dependence on the continuous regressor x. As another example, p(x) may represent the Pr that a car needs warranty service (y =1), but the need of this service may well depend on the accumulated mileage x. Examining Eq. (22a) reveals that if p is a function of a regressor x, then its p^{th} quantile x_p will also be a function of x, say h(x), i.e.,

$$x_p = h(x) = \ln\left[\frac{p(x)}{1-p(x)}\right] = \ln\left[\frac{p(x)}{q(x)}\right] \qquad (22b)$$

In statistical literature, $\frac{p(x)}{1-p(x)}$ is called the odds for success, and ln(Odds) is called the logit transformation (i.e., logit means "$\log_e$ odds of success"). For example, if p(x = 55) represents the Pr that a randomly selected individual has evidence of coronary heart disease (CHD) at the age x = 55, and it is known that

the Odds(55) = $\frac{p(55)}{1-p(55)} = \frac{p(55)}{q(55)}$ = 2.5, then for that randomly-selected person having evidence of CHD at the age of 55 is 2&1/2 times more likely than not having coronary heart disease, i.e., success is 2&1/2 times more likely than failure. Put differently, the Pr of CHD for that 55-year old is 5/7 which is 2&1/2 time q(55) = 2/7. Conversely, if p(x) = 0.80, then 0.80/0.20 = 4 implies that the odds are 4 to 1 in favor of the event of interest. Similarly, if Odds(against an event A) = O(A) are 19 to 1, then its occurrence Pr is 1/(1+19) = 0.05 = $\frac{q(A)}{p(A)}$ = $\frac{19/20}{1/20}$, and it is then said that the odds against the event A are O(A) = 19/1 (or 19 to 1). Rearranging the logit transformation in Eq. (22b) yields

$h(x) = \ln[\frac{p(x)}{1-p(x)}] \rightarrow e^{h(x)} = \frac{p(x)}{1-p(x)} \rightarrow e^{h(x)}[1-p(x)] = p(x) \rightarrow e^{h(x)} = p(x)+p(x)e^{h(x)} \rightarrow$

$1 = p(x)e^{-h(x)} + p(x) \rightarrow$

$$p(x) = \frac{1}{1+e^{-h(x)}} = \frac{e^{h(x)}}{e^{h(x)}+1} = \frac{e^{h(x)}}{1+e^{h(x)}} = [1+e^{-h(x)}]^{-1} \qquad (23)$$

If h(x) is a simple linear regression function, then $h(x) = \beta_0 + \beta_1 x + \in$, and Eq. (23) becomes the simplest form of logit regression function given below.

$$p(x) = Pr(Y=1|\, x) = \frac{e^{\beta_0+\beta_1 x+\in}}{1+e^{\beta_0+\beta_1 x+\in}} = \frac{1}{1+e^{-\beta_0-\beta_1 x-\in}} = \frac{1}{1+e^{-h(x)}} \qquad (24a)$$

Or, combining Eqs. (22b) and (24a) yields the logit (function)

$$h(x) = \ln[\frac{p(x)}{1-p(x)}] = \ln(Odds) = \ln[p(x)] - \ln[q(x)] = \beta_0 + \beta_1 x + \in \qquad (24b)$$

where from Eq. (23) we obtain $q(x) = 1 - p(x) = \frac{1}{1+e^{h(x)}} = [1+e^{h(x)}]^{-1}$.

If h(x) is a function of two regressors, then Eq. (24a) becomes

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\in}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\in}} \qquad (24c)$$

Of course, at this point one could ask the logical question as to why we cannot model p(x) itself in a simple linear regression format, i.e., why can't we let p(x) = $\beta_0 + \beta_1 x + \in$. The problem with such a model is that if we take the conditional expectation of Y, we obtain E(Y|x) = 1×p(x) + 0×[1 – p(x)] = E($\beta_0 + \beta_1 x + \in$) = $\beta_0 + \beta_1 x$ because E($\in$) = 0. This yields E(Y|x) = p(x) = $\beta_0 + \beta_1 x$, where p(x) must lie within the interval [0, 1] while $\beta_0 + \beta_1 x$ cannot be constrained to lie in this interval [0, 1]. Therefore, the link function in logistic regression is the logit h(x) = ln[p(x)/q(x)] = $\beta_0 + \beta_1 x + \in$. If we could model p(x) directly as a simple linear regression of x, then the link would be the identity function. Further, $e^{h(x)}/[1+e^{h(x)}]$ ranges from 0 to 1 just like p(x).

So far we have established how to set up a logistic regression model, but the difficult part is parameter estimation and obtaining of the 95% CIs. To better understand the difficulties, it is best to provide an example with a data set. The following example is borrowed from the text by J. L. Devore's 6th edition (2004) on his page 573 (Duxbury Press), which he refers to as Example 13.6. The data pertain to launch temperature and the failure incidence of O-rings in 24 space flights prior to the Challenger Disaster of January 1986.

Table 8 (Example 13.6 on page 573 of J. L. Devore (2004); X = O-ring Temperature in Fahrenheit)

x	53°F	56	57	63	66	67	67	67	68	69	70	70
Failure	Yes y=1	Yes y=1	Yes y=1	No y=0	No y=0	No y=0	No y=0	No y=0	No y=0	No y=0	No y=0	Yes y=1
°F	70	70	72	73	75	75	76	76	78	79	80	81
Failure	Yes	Yes	No	No	No	Yes	No	No	No	No	No	No

The above table shows that there were a total of 7 failures in n = 24 independent trials, mostly occurring at lower temperatures x. Thus, each launch was a

Bernoulli trial that either failed (y = 1) with probability p(x) or did not fail (y = 0) with probability of 1–p(x) = q(x). Thus, the Bernoulli pmf (Pr Mass Function) of y is given by

$$\text{pmf(y)} = \begin{cases} p(x), & y=1 \\ 1-p(x), & y=0 \end{cases} = \begin{cases} p(x), & y=1 \\ q(x), & y=0 \end{cases} = p(x)^{y} \times q(x)^{1-y}, \quad y = 0 \text{ or } 1 \qquad (25)$$

where x stands for O-ring Temperature, and q(x) = 1 –p(x) represents the Pr that the event of interest does not occur at x, and y = either 0 or 1.

Logistic Regression literature generally suggests three methods of estimation for fitting the occurrence Pr of the event of interest, p(x), to the logistic regression model $\frac{e^{h(x)}}{1+e^{h(x)}} = \frac{e^{\beta_0+\beta_1 x+\epsilon}}{1+e^{\beta_0+\beta_1 x+\epsilon}} = \frac{1}{1+e^{-\beta_0-\beta_1 x-\epsilon}}$: (1) Maximum Likelihood Estimation (MLE), (2) the method of Weighted Least-squares. The 3rd method, discriminant analysis function, is sometimes used but heavily depends on the normality assumption of the independent continuous variable x in each of the two subgroups (0 & 1), and moderate departures of X from normality leads to very positively biased estimators of β_0 & β_1. Further, the most commonly used estimation procedure in Logistic Regression literature is the MLE.

Maximum Likelihood Estimation (MLE)

Consider the regression logistic model

$$p(x) = \frac{e^{\beta_0+\beta_1 x+\epsilon}}{1+e^{\beta_0+\beta_1 x+\epsilon}} = \frac{1}{1+e^{-h(x)}} \qquad (26)$$

As illustrated in Table 8, we have n = 24 pairs of (x_i, y_i) = (53, 1), (56, 1), (57, 1), (63, 0), …, (80, 0), (81, 0). The likelihood function (LF) for Eq. (26) and Table 8 is given by

$$L(\beta_0, \beta_1) = p(53)\times p(56)\times p(57)\times q(63)\times q(66) \times q(67)\ldots \times q(80)\times q(81).$$

$$= \prod_{i=1}^{n=24} p(x_i)^{y_i} \times q(x_i)^{1-y_i}, \quad y_i = 0 \text{ or } 1. \qquad (27a)$$

where we have made use of Eq. (25) and the fact that the 24 Bernoulli trials are

independent. Because maximizing the log of a function also maximizes the function itself, it will be easiest to take the natural log of the LF in (27a) and attempt to maximize it. Further to simplify notation, we let $p_i = p(x_i)$ and $q_i = 1 - p(x_i)$, and clearly the corresponding logit is $h(x_i) = \beta_0 + \beta_1 x_i + \in_i$.

$$L(\beta_0, \beta_1) = \ln[L(\beta_0, \beta_1)] = \ln \prod_{i=1}^{n} p(x_i)^{y} \times q(x_i)^{1-y} = \sum_{i=1}^{n} \left[\ln p(x_i)^{y_i} + \ln q(x_i)^{1-y_i} \right]$$

$$= \sum_{i=1}^{n} \left[y_i \ln p_i + (1 - y_i) \ln q_i \right] = \sum_{i=1}^{n} \left[y_i (\ln p_i - \ln q_i) + \ln q_i \right]$$

$$= \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) + \ln(\frac{1}{1 + e^{h(x_i)}})] = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - \ln(1 + e^{h(x_i)})] \qquad \textbf{(27b)}$$

where I have made use of Eqs. (24 a & b), leaving out $\in_i$ because partial differentiation will not be affected by $\in_i$. We first differentiate (27b) wrt β_0:

$$\partial L(\beta_0, \beta_1)/\partial \beta_0 = \sum_{i=1}^{n} [y_i(1) - \frac{\partial}{\partial \beta_0} \ln(1 + e^{h(x_i)})] = \sum_{i=1}^{n} [y_i - \frac{1}{1 + e^{h(x_i)}} \frac{\partial}{\partial \beta_0} (1 + e^{h(x_i)})]$$

$$= \sum_{i=1}^{n} [y_i - \frac{e^{h(x_i)}}{1 + e^{h(x_i)}} \frac{\partial}{\partial \beta_0} h(x_i)] = \sum_{i=1}^{n} [y_i - p(x_i) \times (1)] = \sum_{i=1}^{n} [y_i - p(x_i)]$$

$\xrightarrow{\text{Set equal to}} 0 \rightarrow$ Thus, our first likelihood equation is $\sum_{i=1}^{n} \hat{p}(x_i) = \sum_{i=1}^{n} y_i \rightarrow$ this implies that the sum of observed values of y must equal to the sum of the fitted values $\hat{p}(x)$ as in the case of classical MLREG. For Table 8, we obtain

$$\sum_{i=1}^{n} \hat{p}(x_i) = \sum_{i=1}^{n} [\frac{1}{1 + e^{-\hat{\beta}_0 - \hat{\beta}_1 x_i}}] = \sum_{i=1}^{n} [\frac{1}{1 + e^{-\hat{h}(x_i)}}] = \sum_{i=1}^{n} [1 + e^{-\hat{h}(x_i)}]^{-1} = 7.0 = n_1$$

where n_1 represents the number of O-ring failures in n = 24 Bernoulli trials, and $n_0 = n - n_1 = 17$.

We next differentiate the log-likelihood function in Eq. (27b) wrt β_1:

$$\partial L(\beta_0, \beta_1)/\partial \beta_1 = \sum_{i=1}^{n} [y_i x_i - \frac{\partial}{\partial \beta_1} \ln(1 + e^{h(x_i)})] = \sum_{i=1}^{n} [y_i x_i - \frac{e^{h(x_i)}}{1 + e^{h(x_i)}} \frac{\partial}{\partial \beta_i} h(x_i)] =$$

$\sum_{i=1}^{n}[y_i x_i - p(x_i)\times(x_i)] \xrightarrow{\text{Set equal to}} 0$

Thus our 2nd likelihood equation becomes $\sum_{i=1}^{n} x_i \hat{p}(x_i) = \sum_{i=1}^{n} y_i x_i$: for Table 8

we have: $\sum_{i=1}^{n}[\frac{x_i}{1+e^{-\hat{\beta}_0-\hat{\beta}_1 x_i}}] = \sum_{i=1}^{24} x_i y_i$ = 53 + 56 + 57 + 0 + …+ 75 = 451.

I used the Excel solver and obtained the point estimates $\hat{\beta}_0$ = 10.87535 and $\hat{\beta}_1$ = – 0.17132 $\rightarrow \hat{h}(x) = 10.87535 - 0.17132x$; Minitab (Enter x and y pairs in C1 & C2 and go to Stat → Regression → Binary Logistic Regression and the rest should be self-explanatory) also verified my answers. Thus, $\ln[\frac{\hat{p}(x)}{1-\hat{p}(x)}] = \ln[\frac{\hat{p}(x)}{\hat{q}(x)}] =$ $\ln(\widehat{Odds}) = \hat{\beta}_0 + \hat{\beta}_1 x = \hat{h}(x) = 10.87535 - 0.17132x$; for example, when temperature is equal to 60°F, then $\ln[\widehat{O}dds\,(60°F)] = 10.87535 - 0.17132(60) = 0.596119 \rightarrow$ $\widehat{O}dds\ \text{of success}\,(60°F) = e^{0.596119} = 1.815061 \rightarrow$ then at 60 degrees Fahrenheit O-ring failure is 1.8151 times more likely than no failure, while at 61°F, $\ln[\widehat{O}dds\,(61°F)] =$ 0.42483 $\rightarrow \widehat{O}dds\,(61°F) = 1.5293304114 \rightarrow$ this implies that for every one degree increase in temperature (°F) the odds of failure diminishes by a factor of 1.5293304114/1.815061 = 0.8425515. Note that Minitab reports a value of 0.84 for this last factor that is called the Odds Ratio (OR). That is, for every unit increase in x, the OR diminishes by roughly 0.8426 because $\widehat{OR} = \widehat{O}dds(x+1)/\widehat{O}dds(x) =$ $e^{\hat{h}(x+1)}/e^{\hat{h}(x)} = e^{\hat{\beta}_0+\hat{\beta}_1(x+1)}/e^{\hat{\beta}_0+\hat{\beta}_1 x} = e^{\hat{\beta}_1} = e^{-0.17132} = 0.8425515$. Thus, in general a one unit increase in x results in the estimated odds $\hat{p}(x)/\hat{q}(x)$ multiplied by a factor of $e^{\hat{\beta}_1}$. Because $\hat{\beta}_1$ < 0, the relationship between p(x) and x is a decreasing one. The resulting logistic regression function is given by: $\hat{p}\,(x) = \frac{1}{1+e^{-10.87535+0.17132x}}$. This last logistic model implies that when x = 60°F, the O-ring failure Pr is estimated as $\hat{p}(60) = [1+e^{-10.87535+0.17132(60)}]^{-1}$ = 0.64477, while at x

= 61°F, $\hat{p}(61) = 0.60463$. Note that the value of the odds ratio is also given by OR = $(0.60463 \times 0.35523)/[\hat{q}(61) \times \hat{p}(60)] = 0.8425515$, i.e., OR = $\dfrac{\hat{p}(x+1)/\hat{q}(x+1)}{\hat{p}(x)/\hat{q}(x)}$.

Computing the *se*'s of the Maximum Likelihood Estimators

It has been proven in statistical theory that the covariance matrix of the ML estimators (MLEs) asymptotically approaches the inverse of the Fisher's information matrix given below

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \qquad (28)$$

where $I_{ij} = E\left[-\dfrac{\partial^2 L}{\partial\beta_i \partial\beta_j}\right] = I_{ji}$, L represents the natural log of the likelihood function and is given in Eq. (27b) for the case of two parameters β_0 and β_1. Further, from statistical theory, maximum likelihood estimators are generally biased but are asymptotically (as $n \to \infty$) unbiased. We now proceed to compute the *se*'s of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the data of Table 8.

$$\frac{\partial^2 L}{\partial\beta_0^2} = \frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}[y_i - p(x_i)] = -\sum_{i=1}^{n}\frac{\partial}{\partial\beta_0}p(x_i) = -\sum_{i=1}^{n}\frac{\partial}{\partial\beta_0}[1+e^{-h(x_i)}]^{-1} =$$

$$-\sum_{i=1}^{n}-[1+e^{-h(x_i)}]^{-2}\frac{\partial}{\partial\beta_0}(1+e^{-\beta_0-\beta_1 x_i}) = \sum_{i=1}^{n}[1+e^{-h(x_i)}]^{-2}[-e^{-h(x_i)}] = -\sum_{i=1}^{n}p^2(x_i)e^{-h(x_i)} \cong$$

-3.73459135; similarly,

$$\frac{\partial^2 L}{\partial\beta_0\partial\beta_1} = -\sum_{i=1}^{n}x_i p^2(x_i)e^{-h(x_i)} \cong -254.20087919, \text{ and}$$

$$\frac{\partial^2 L}{\partial\beta_1^2} = -\sum_{i=1}^{n}x_i^2 p^2(x_i)e^{-h(x_i)} \cong -17446.21098707.$$ Note that taking the expectation of $\dfrac{\partial^2 L}{\partial\beta_i\partial\beta_j}$ is generally mathematically intractable, at least to the ability of this author, and thus I have approximated them by their sample values.

Further, $\frac{\partial^2 L}{\partial \beta_0^2} < 0$, $\frac{\partial^2 L}{\partial \beta_1^2} < 0$ and $\frac{\partial^2 L}{\partial \beta_0^2} \times \frac{\partial^2 L}{\partial \beta_1^2} > (\frac{\partial^2 L}{\partial \beta_0 \partial \beta_1})^2$, implying that the response surface is strictly concave and hence, a global maximum.

Thus, the fisher's information matrix is estimated by (recall that $I_{ij} = E\left[-\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}\right]$)

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = \begin{bmatrix} 3.73459 & 254.20088 \\ 254.20088 & 17446.21099 \end{bmatrix} \rightarrow I^{-1} = \begin{bmatrix} 32.52574 & -0.473918 \\ -0.473918 & 0.00696256 \end{bmatrix}$$

$\rightarrow se(\hat{\beta}_0) = \sqrt{32.52574} = 5.70313434$ and $se(\hat{\beta}_1) = \sqrt{0.00696256} = 0.083442$. These standard errors precisely match those of Minitab's. Therefore, the 95% Confidence limits for β_0 are $\hat{\beta}_0 \mp z_{0.025} \times se(\hat{\beta}_0) = 10.87535 \mp 1.96 \times 5.703134 =$ (– 0.30279330, 22.0534933) $\rightarrow -0.30279330 \leq \beta_0 \leq 22.0534933$

Similarly, the 95% CI for β_1 is given by $\hat{\beta}_1 \mp z_{0.025} \times se(\hat{\beta}_1) \rightarrow$

$$-0.33486678 \leq \beta_1 \leq -0.00777422$$

Note that in developing the above CI's, I have used the fact that all ML estimators in the universe are asymptotically normally distributed. A larger sample size n leads to a more accurate normal approximation. The CI for the Odds ratio is obtained from $OR_L = e^{-0.33486678} = 0.7154334$ and $OR_U = e^{-0.00777422} = 0.992256 \rightarrow 0.7154334 \leq OR \leq 0.992256$. This last 95% CI also agrees with that of Minitab's to 2 decimals.

The Method of Weighted Least-Squares

This method generally requires repeated observations at different levels of x so that a frequency distribution can be constructed. [To illustrate the procedure, we use the data from the text by D. W. Hosmer and S. Lemeshow (*Applied Logistic Regression*, 2nd Ed., wiley, ISBN:0-471-35632-8) in their Table 1.1 on page 3 where the variable x represents the age of an individual in the study and y represents the presence or absence of CHD (Coronary Heart Disease)]. Because there were 100 individuals in their sample, the authors proceeded to obtain a frequency distribution for their data which is, except for minor modifications by me, duplicated atop the next page. Note that in each age

Table 1.2 on page 3 of Hosmer & Lemeshow (2000)

Age Group	n_i	CHD Present	$\hat{p}_i$	$\hat{q}_i$	$\hat{w}_i = n_i\hat{p}_i\hat{q}_i$
20-29 years	10	1	0.10	0.90	0.90
30-34	15	2	2/15	13/15	1.733333
35-39	12	3	0.25	0.75	2.25
40-44	15	5	1/3	2/3	3.33333
45-49	13	6	6/13	7/13	3.23077
50-54	8	5	5/8	3/8	1.875
55-59	17	13	13/17	4/17	3.05882
60-69	10	8	0.80	0.20	1.600
Totals	100	43	0.43	0.57	17.9813

subgroup we have a binomial random variable with n_i Bernoulli trials and estimated success Pr of $\hat{p}_i$. The likelihood function for observing n_1 successes in n trials is given by $L(y_1, y_2, \ldots, y_n; p) = \frac{n!}{n_1!(n-n_1)!}p^{n_1}(1-p)^{n-n_1} = Cp^{n_1}q^{n-n_1}$, where $C = n!//[n_1!(n-n_1)!] = {}_nC_{n-n_1}$ is a constant (i.e., free of p) and q = 1–p.

$$L(p) = \ln\left[\frac{n!}{n_1!(n-n_1)!}p^{n_1}(1-p)^{n-n_1}\right] = \ln(C) + n_1\ln(p) + (n-n_1)\ln(q)$$

Setting $dL(p)/dp$ equal to zero results in the ML estimate of p as $\hat{p} = n_1/n$.

Secondly, it can be shown that $d^2L(p)/dp^2 = -n/(pq)$ so that the $V(\hat{p}) = pq/n$.

Because the variance of the Binomial rv is equal to npq, the estimated variance at each subgroup is $n_i\hat{p}_i\hat{q}_i$. We now regress ln(Odds) versus age x.

$\ln(\text{Odds for CHD}) = \ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{p_i}{q_i}\right) = \beta_0 + \beta_1 x_i + \in_i = \text{logit}(x_i)$. In weighted regression, we try to give less weight to the subgroup with the larger variance and assume that $V(y|x_i)$ depends on the level of x through $V(y|x_i) = \sigma^2_\in/w_i$, i.e., the weighted LSF is given by

$$\text{WLSF}(\beta_0, \beta_1) = \sum_{i=1}^{k} \epsilon_i^2 = \sum_{i=1}^{k} w_i (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (29)$$

where $y_i = \ln(\frac{p_i}{1-p_i})$ and k represents the number of categories (or subgroups) and is equal to 8 for Table 1.2 of Hosmer & Lemeshow reproduced above and $y_i = \beta_0 + \beta_1 x_i + (\epsilon_i / \sqrt{w_i})$ so that $V(y|x_i) = \sigma_\epsilon^2 / w_i$. Differentiating Eq. (29) wrt to β_0 and β_1 and setting the resulting derivatives equal to zero leads to the following system of weighted normal equations.

$$\hat{\beta}_0 \sum_{i=1}^{k} w_i + \hat{\beta}_1 \sum_{i=1}^{k} w_i x_i = \sum_{i=1}^{k} w_i y_i \qquad (30a)$$

$$\hat{\beta}_0 \sum_{i=1}^{k} w_i x_i + \hat{\beta}_1 \sum_{i=1}^{k} w_i x_i^2 = \sum_{i=1}^{k} w_i x_i y_i \qquad (30b)$$

Letting $W = \sum_{i=1}^{k} w_i$ and solving the 1st normal equation (30a) for $\hat{\beta}_0$ gives

$$\hat{\beta}_0 = (\sum_{i=1}^{k} w_i y_i - \hat{\beta}_1 \sum_{i=1}^{k} w_i x_i)/W = \bar{y} - \hat{\beta}_1 \bar{x} \quad (\text{where } \bar{y} = \Sigma w_i y_i / W)$$

Substituting the above expression for $\hat{\beta}_0$ into (30b) results in

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{k} w_i x_i y_i - (\sum_{i=1}^{k} w_i x_i)(\sum_{i=1}^{k} w_i y_i)/W}{\sum_{i=1}^{k} w_i x_i^2 - (\sum_{i=1}^{k} w_i x_i)^2 / W} = \frac{\sum_{i=1}^{k} w_i (y_i - \bar{y}) x_i}{\sum_{i=1}^{k} w_i (x_i - \bar{x})^2}$$

Using the Hosmer & Lemeshow's (2000) grouped data in their Table 1.2, I obtained $W = 17.98126$, $\sum_{i=1}^{8} w_i x_i = 827.6658$ (using subgroup midpoints for x_i's), $\sum_{i=1}^{8} w_i y_i = -3.72119$, where $y_i = \ln(\frac{\hat{p}_i}{\hat{q}_i})$, $\sum_{i=1}^{k=8} w_i x_i y_i = 34.19435$, $\sum_{i=1}^{8} w_i x_i^2 = 40076.7$, $\hat{\beta}_1 = 0.103789$ and $\hat{\beta}_0 = -4.98427$. Hosmer & Lemeshow used the ML procedure

and give the ML estimates as $\hat{\beta}_1 = 0.111$ and $\hat{\beta}_0 = -5.309$. Note that as n increases, the values of weighted least-squares estimators approach those of the ML. Thus,

$$\ln(\text{Odds for CHD}) = \ln\left(\frac{\hat{p}_i}{\hat{q}_i}\right) = -4.98427 + 0.103789x_i$$

and $\hat{p}(x_i) = (1+e^{4.98427-0.103789x_i})^{-1} \rightarrow \hat{p}(50) = 0.551111 \rightarrow$ Thus, a randomly selected 50 year-old person has 0.5511 estimated Pr of having evidence of CHD. Using Hosmer & Lemeshow ML estimates, the same Pr at age 50 is equal to 0.55996.

Because ML estimates have nicer properties [1: they are asymptotically unbiased, (2) their SMD approaches normality as $n \rightarrow \infty$, (3) if $\hat{\theta}$ is the ML estimator of a parameter θ, then $h(\hat{\theta})$ is the ML estimator of $h(\theta)$], I used the raw data in Table 1.1 of Hosmer & Lemeshow (2000) and proceeded to obtain their ML estimates. As before, the two likelihood equations with two unknowns are

$$\sum_{i=1}^{n} \hat{p}(x_i) = \sum_{i=1}^{n}\left[\frac{1}{1+e^{-\hat{\beta}_0-\hat{\beta}_1 x_i}}\right] = \sum_{i=1}^{n}[1+e^{-\hat{h}(x_i)}]^{-1} = 43 = \sum_{i=1}^{n} y_i = n_1$$

$$\sum_{i=1}^{n} x_i\hat{p}(x_i) = \sum_{i=1}^{n} y_i x_i \rightarrow \sum_{i=1}^{n}\left[\frac{x_i}{1+e^{-\hat{\beta}_0-\hat{\beta}_1 x_i}}\right] = \sum_{i=1}^{n} x_i(1+e^{-\hat{\beta}_0-\hat{\beta}_1 x_i})^{-1} = 2205.$$

The Excel file on my website provides their raw data. The use of Excel solver yielded the solution set $\hat{\beta}_0 = -5.30945332$ and $\hat{\beta}_1 = 0.11092114$ (as compared to the reported values of -5.309 & 0.111 on their page 10). Thus,

$$\hat{h}(x) = -5.30945332 + 0.11092114x$$

Note that Hosmer & Lemeshow (2000) use g(x) to represent $\beta_0 + \beta_1 x + \in$, but I have used g(x) to denote the pdf of a Logistic random variable. A point estimate of the logit at 50 years of age is given by $\hat{h}(50) = 0.236603734$ so that a point estimate of the odds is $\hat{O}dds(\text{at age 50 for CHD}) = e^{0.236603734} = 1.266938973 \rightarrow$ this implies that a random 50-year old person is 1.267 times more likely to have evidence of CHD than not to have evidence of CHD. Further, the point estimate

of logistic Pr at age x = 50 is $\hat{p}(50)$ = 1/(1+e $^{-0.236603734}$) = 0.558876524, or $\hat{p}(50)$ = 1.266938973/(1+1.266938973) = 0.558876524 and $\hat{q}(50)$ = 0.441123476.

The corresponding Fisher's information matrix is

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = \begin{bmatrix} 17.92631205 & 826.1196216 \\ 826.1196216 & 39798.5425534 \end{bmatrix} \rightarrow I^{-1} =$$

$$\begin{bmatrix} 1.285173 & -0.02667702 \\ -0.02667702 & 0.0005788757 \end{bmatrix} \rightarrow se(\hat{\beta}_0) = \sqrt{1.285173} = 1.133655 \text{ and}$$

$se(\hat{\beta}_1) = \sqrt{0.0005788757} = 0.02406$. These standard errors precisely match those of the authors to 4 decimals.

95% CI for the Fitted Logistic Regression p(x)

Because $p(x) = \dfrac{e^{h(x)}}{1+e^{h(x)}} = \dfrac{e^{\beta_0+\beta_1 x+\epsilon}}{1+e^{\beta_0+\beta_1 x+\epsilon}} = [1+e^{-h(x)}]^{-1}$, then we first need to obtain the requisite CI for h(x) and this in turn will give us the 95% CI for p(x). Therefore, we need to compute the *se* of the estimated logit $\hat{h}(x)$, which requires computing the V[$\hat{h}(x)$] first followed by taking its square root. Proceeding with applying the variance operator to $\hat{h}(x)$, we obtain

$$V[\hat{h}(x)] = V(\hat{\beta}_0 + \hat{\beta}_1 x) = V(\hat{\beta}_0) + x^2 V(\hat{\beta}_1) + 2x[Cov(\hat{\beta}_0, \hat{\beta}_1)] \qquad (31)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are correlated estimators because the information matrix in Eq. (28) is not a diagonal matrix. The last term on the RHS of Eq. (31) gives the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ which is defined as

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = E\{[(\hat{\beta}_0 - E(\hat{\beta}_0)]\times[\hat{\beta}_1 - E(\hat{\beta}_1)]\} = E(\hat{\beta}_0 \hat{\beta}_1) - E(\hat{\beta}_0)\times E(\hat{\beta}_1)$$

Note that all ML estimators are only asymptotically unbiased, i.e., E($\hat{\beta}$) is in general different from the parameter β for n $\leq$ 50.

For the Example by Hosmer & Lemeshow (their Table 1.1), we have

$I^{-1} = \begin{bmatrix} 1.285173 & -0.02667702 \\ -0.02667702 & 0.0005788757 \end{bmatrix}$, which shows from Eq. (31) that

$$\hat{V}[\hat{h}(x)] = 1.285173 + x^2(0.0005788757) + 2x(-0.02667702) \qquad (32)$$

Thus, for a person of age x = 50, Eq. (32) gives $\hat{V}[\hat{h}(50)]$ = 1.285173 + $(50)^2(0.0005788757) + 100(-0.02667702) = 0.06466025 \rightarrow se[\hat{h}(50)] = 0.2542838 \rightarrow$ $h_L(50) = 0.236603680 - 1.96 \times 0.2542838 = -0.2617925644$, and $h_U(50) = 0.236603680 + 1.96 \times 0.2542838 = 0.7350$. Thus, the lower 95% confidence limit for the fitted p(x) at age 50 is $p_L(50) = (1+e^{0.2617926})^{-1} = 0.43492311$ and $p_U(50) = (1+e^{-0.735})^{-1} = 0.67590153 \rightarrow 0.434923 \leq p(50) \leq 0.675902 \rightarrow$ the Pr that a randomly-selected 50-year old person will have evidence of CHD, before that person is selected, lies within the random interval $[p_L(50), p_U(50)]$ is 95%, where the CI end points and length change for one random sample of size 100 to the next. Put differently, in repeated sampling, say 100000 samples each of size n =100, roughly 95000 of the CIs (of differing end points) will contain the true mean proportion p(50).

The Likelihood Ratio Statistic for Testing $H_0 : \beta_1 = 0$

In the field of statistics, the likelihood ratio statistic for testing $H_0: \beta_1 = 0$ is defined as

$$\text{LRS} = \frac{\text{Max } L(x|H_0)}{L(x|\hat{\beta})} \qquad (33a)$$

where $\hat{\beta} = [\hat{\beta}_0 \quad \hat{\beta}_1]'$ for the case of simple LREG. Some authors in statistical literature use λ for LRS and others denote LRS by Λ. For simplicity, I will just use LRS to denote the likelihood ratio statistic. The denominator of LRS, $L(x|\hat{\beta})$ is the value of the likelihood function [see Eq. (33a)] when all the parameters in the density function are replaced by their corresponding ML estimates, while the numerator is the maximum of LF wrt only some of the parameters while the remaining parameters are restricted under H_0. Because the denominator of (33a) maximizes $L(x|\beta)$ wrt to all parameters while the numerator wrt to only some of the parameters, then the numerator can never exceed the denominator, and hence, the likelihood ratio statistic is restricted to the interval $0 \leq \text{LRS} \leq 1$. Thus for our simple logistic regression of testing $H_0 : \beta_1 = 0$, the LRS reduces to

$$\text{LRS} = \frac{L(x|\hat{\beta}_0)}{L(x|\hat{\beta}_0,\hat{\beta}_1)} \qquad (33b)$$

Note that the numerator assumes that β_1 is set equal to zero as stated under H_0. Further, when the values of numerator and dominator of (33b) are close to each other, then the value of $\hat{\beta}_1$ must be close to zero in agreement with H_0. However, when the denominator is much larger than numerator, then $\hat{\beta}_1$ must be far from zero leading to the rejection of H_0: $\beta_1 = 0$. Therefore, the 5% rejection region for testing H_0: $\beta_1 = 0$ corresponds to small values of LRS having a Pr of at most 0.05. Fortunately, although the exact sampling distribution (SMD) of LRS in (33b) is difficult to obtain and intractable when the underlying distribution is unknown, from statistical theory the SMD of $-2\times\ln(\text{LRS})$ approaches a χ^2_ν as $n \to \infty$, where the degrees of freedom ν is equal to the number of parameters hypothesized under H_0 (in this case $\nu = 1$). That is, $-2\times\ln(\text{LRS}) \to \chi^2_1$ as n increases towards infinity for testing H_0: $\beta_1 = 0$. For the sake of illustration, I will compute $-2\times\ln(\text{LRS}) = -2\times\ln\left[\frac{L(x|\hat{\beta}_0)}{L(x|\hat{\beta}_0,\hat{\beta}_1)}\right] = -2\times\ln[L(x|\hat{\beta}_0)] + 2\times\ln[L(x|\hat{\beta}_0,\hat{\beta}_1)] = D_0 - D_1$, which Hosmer & Lemeshow (2000) denote it by G, i.e.,

$$G = -2\times\ln(\text{LRS}) = -2\times\ln\left[\frac{L(x|\hat{\beta}_0)}{L(x|\hat{\beta}_0,\hat{\beta}_1)}\right] = D_0 - D_1 \qquad (34)$$

where $D_0 = -2\times\ln[L(x|\hat{\beta}_0)]$ is called the deviance with $\beta_1 = 0$, $D_1 = -2\times\ln[L(x|\hat{\beta}_0,\hat{\beta}_1)]$ is the deviance with $\beta_1 \neq 0$, and G has an approximate χ^2 distribution with $\nu = 1$ df. For the data of Table 1.1 of Hosmer & Lemeshow (listed on my website) the value of D_1 from Eq. (33b) is given by

$$D_1 = -2\times\ln[L(x|\hat{\beta}_0,\hat{\beta}_1)] = -2\times\sum_{i=1}^{n}[y_i\ln\hat{p}_i + (1-y_i)\ln\hat{q}_i] = -2\times(-53.6765463) = 107.3530927.$$

In order to compute D_0, we let $h(x) = \beta_0 + \in$, i.e., we assume H_0: $\beta_1 = 0$ is true. $\rightarrow$

$L(\beta_0, \beta_1 = 0) = \sum_{i=1}^{n} [y_i \beta_0 - \ln(1 + e^{\beta_0})] \rightarrow \partial L(\beta_0)/\partial\beta_0 = \sum_{i=1}^{n} [y_i - \frac{e^{\beta_0}}{1 + e^{\beta_0}})]$

$\xrightarrow{\text{Set Equal to}} 0 \rightarrow \hat{\beta}_0 \,|\, H_0 \text{ is true} = \ln(\frac{\bar{y}}{1 - \bar{y}}) = \ln(0.43/0.57) = -0.281851152$

$\rightarrow L(\hat{\beta}_0 \,|\, H_0 : \beta_1 = 0 \text{ is true}) = \sum_{i=1}^{n} [y_i \hat{\beta}_0 - \ln(1 + e^{\hat{\beta}_0})] = \hat{\beta}_0 \sum_{i=1}^{n} y_i - n \times \ln(1 + e^{\hat{\beta}_0})$

$= -68.3314914 \rightarrow D_0 = -2 \times \ln[L(x \,|\, \hat{\beta}_0)] = 136.662983 \rightarrow G = -2 \times \ln(LRS) =$

136.662983 – 107.3530927 = 29.30989. This compares exactly to 2 decimals with Hosmer & Lemeshow's answer of G = 29.31 atop their page 15. Lastly, we compute the Pr level for testing H_0: $\beta_1 = 0$, which is given by $\hat{\alpha}$ = *P*-value= *p* = $\Pr(\chi_1^2 \geq 29.30989) = 0.0000000616801 = 0.0^7616801 \rightarrow$ Very strongly reject H_0: $\beta_1 = 0 \rightarrow$ the age of a person is a strong predictor of evidence of CHD.

The Score Test (ST) Statistic for Testing H_0: $\beta_1 = 0$

Besides the Z-statistic, $\hat{\beta}_1 / se(\hat{\beta}_1)$, the LRS, and G, there is one other statistic, called the Score Test (ST), for testing H_0: $\beta_1 = 0$. This is based on the value of $\partial L(\beta_0, \beta_1)/\partial\beta_1$ given that H_0: $\beta_1 = 0$ and $\partial L(\beta_0, \beta_1)/\partial\beta_0 = 0$ are true. Recall that the ML estimate of β_1 is obtained by setting both $\partial L(\beta_0, \beta_1)/\partial\beta_0$ and $\partial L(\beta_0, \beta_1)/\partial\beta_1$ equal to zero and solving the resulting system of the two likelihood equations for the ML estimates of β_0 and β_1. Under the null hypothesis H_0: $\beta_1 = 0$, the 1st likelihood equation $\partial L(\beta_0, \beta_1)/\partial\beta_0 = 0$ yields $\sum_{i=1}^{n} \hat{p}(x_i) = \sum_{i=1}^{n} y_i = n_1 \rightarrow$

$\sum_{i=1}^{n} [\frac{1}{1 + e^{-\hat{\beta}_0}}] = \sum_{i=1}^{n} y_i \rightarrow n[\frac{1}{1 + e^{-\hat{\beta}_0}}] = \sum_{i=1}^{n} y_i \rightarrow 1 + e^{-\hat{\beta}_0} \rightarrow 1/\bar{y} \rightarrow e^{-\hat{\beta}_0} = -1 +$

$1/\bar{y} \rightarrow -\hat{\beta}_0 = \ln(\frac{1-\bar{y}}{\bar{y}}) \rightarrow \hat{\beta}_0 = \ln(\frac{\bar{y}}{1-\bar{y}}) = \ln(\frac{n_1/n}{1-n_1/n}) \rightarrow \hat{\beta}_0 = \ln(\frac{n_1}{n-n_1}) =$

$\ln(\frac{n_1}{n_0})$, $\hat{p}(x_i) = \bar{y} = n_1/n$, and $n_0 = n - n_1$.

Because the ML estimate of β_1 is obtained from $\partial L(\beta_0, \beta_1)/\partial\beta_1 = 0$, and thus, if $\partial L(\beta_0, \beta_1)/\partial\beta_1$ is far away from zero, its large distance from zero will be in contradiction with the null hypothesis H_0: $\beta_1 = 0$. Therefore, we must compute $\partial L(\beta_0, \beta_1)/\partial\beta_1$ assuming that H_0 and $\partial L(\beta_0, \beta_1)/\partial\beta_0 = 0$ are true and assess if it is significantly different from zero. Assuming H_0 is true, we obtain $\partial L(\beta_0, \beta_1)/\partial\beta_1 = \sum_{i=1}^{n}[y_i x_i - x_i\hat{p}(x_i)] = \sum_{i=1}^{n}[y_i x_i - \bar{y}\times x_i] = \sum_{i=1}^{n} x_i(y_i - \bar{y})$. It is assumed that this last partial derivative is approximately Gaussian with mean zero under H_0 and conditional variance, keeping x_i fixed, that is computed below.

$V[\sum_{i=1}^{n} x_i(y_i - \bar{y})] = V[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})] = V[\sum_{i=1}^{n}(x_i - \bar{x})y_i]$

$= \sum_{i=1}^{n}(x_i - \bar{x})^2 V(y_i) = \sum_{i=1}^{n}(x_i - \bar{x})^2 pq = S_{xx}\times pq \rightarrow$ Thus, the estimate of the

$V[\sum_{i=1}^{n} x_i(y_i - \bar{y})]$ is $\sum_{i=1}^{n}(x_i - \bar{x})^2\hat{p}\hat{q} = S_{xx}\times\bar{y}(1-\bar{y})$. As a result, the Score Test Statistic is given by $ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})S_{xx}}}$, which is approximately normally distributed with zero mean and STDEV 1. For Table 1.1 of Hosmer and Lemeshow (2000), we have: $\bar{y} = 43/100 = 0.43$, $1 - \bar{y} = 0.57$, $S_{xx} = 210560 - 4438^2/100 = 13601.5600$, $\sum_{i=1}^{n} x_i y_i = 2205$, $\bar{y}\sum_{i=1}^{n} x_i = 0.43\times 4438 = 1908.34 \rightarrow$

$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = 2205 - 1908.34 = 296.66 \rightarrow ST = \frac{296.66}{\sqrt{0.43\times 0.57\times 13601.56}} = 5.137987$

$\rightarrow$ *P*-value = Pr Level of the Test = $\hat{\alpha} = 2\Pr(Z_{N(0,1)} \geq 5.137987) = 2.78198\times 10^{-7} \rightarrow$

Strongly reject H_0: $\beta_1 = 0$. Note that $ST^2 = 5.137987^2 = 26.39891$ is fairly close to G = 29.30989, as expected because $Z^2_{N(0,1)} \sim \chi^2_1$ distributed. The discrepancy is due to the fact that G = –2×ln(LRS) is only approximately χ^2_1 distributed.

Goodness-of-Fit (GOF) Tests in Logistic Regression

Minitab reports three GOF statistics to ascertain how well the logistic model fits a binary data set. (1) Pearson Chi-Square Statistic, (2) Deviance, and (3) Hosmer-Lemeshow Test. The last (H&L) is a contingency type of GOF test using 10 subgroups (based on 10 deciles) and in my opinion should not be used unless n > 50. Even, for 50 < n ≤ 100, some of the expected frequencies may turn out to be less than 5 and thus the SMD of the GOF statistic $\sum_{i=1}^{10} \frac{(f_j - E_j)^2}{E_j}$ would not closely resemble χ^2_8, where there are k = 10 subgroups with two constraints $\partial L(\beta_0, \beta_1)/\partial\beta_0 = 0$ and $\partial L(\beta_0, \beta_1)/\partial\beta_1 = 0$. Therefore, I will discuss only the first two GOF tests, starting with Deviance.

Recall that the observed values of y can equal to either 1 (when success occurs) or zero (when the event of interest does not occur). When y = 1 at an x that is not repeated, then the fit is excellent iff $\hat{p}(x)$ is close to 1 because p(x) = Pr(Y =1| x), and thus, $\ln[1/\hat{p}(x)]$ is a measure of goodness-of-fit of the logistic regression model at y = 1 because a large value of $\ln[1/\hat{p}(x)]$ implies that $\hat{p}(x)$ must be close to zero and the model does not fit y = 1 at this x-value. Thus, the y = 1 deviance residual is defined as $d[1, \hat{p}(x)] = \{2\times\ln[1/\hat{p}(x)]\}^{1/2} = \sqrt{-2\times\ln[\hat{p}(x)]}$ iff x is not repeated. Similarly, when y = 0 at an x that is not repeated, the fit is excellent iff $\hat{q}(x)$ is close to 1, but the case y = 0 creates the problem that $\ln[(0/\hat{q}(x)]$ is not defined; further, we have to define the y = 0 deviance residual in such a manner that it is always negative because the y = 1

deviance residual is always positive when x is not repeated. For this reason, the y = 0 logistic residual is basically defined as $\ln[\hat{q}(x)] < 0 \rightarrow 2\times\ln[\hat{q}(x)] < 0 \rightarrow \sqrt{2\times\ln[\hat{q}(x)]}$ is imaginary, and hence, the y = 0 deviance residual is defined as $d[0, \hat{q}(x)] = -\sqrt{-2\times\ln\hat{q}(x)}$. Note that when y = 0 at an x that is distinct from all other subjects, a negative value of $\ln[\hat{q}(x)]$ close to zero is a measure of goodness-of-fit. Thus, for the portion of the data where x is not repeated the total deviance residual is given by $DR_1 = \sum_{\text{All x's are different}} \{d[1, \hat{p}(x)] + d[0, \hat{q}(x)]\}$. If there are at least two subjects with the same value of x_j, there cannot be but one fitted value $\hat{p}_j$ for all the subjects with the same x_j. For example, for the Challenger data on my website, x = 70°F is replicated 4 times with the results y = 1 three times and y = 0 once but with the same logistic Pr equal to $\hat{p}(70) = 0.24655$. For such a case, the deviance residual is a weighed average of the $d[1, \hat{p}(x)]$ and $d[0, \hat{q}(x)]$ given below:

$$d(y_j, p_j) = \pm\sqrt{2\left[y_j \ln\left(\frac{y_j}{n_j p_j}\right) + (n_j - y_j)\ln\left(\frac{n_j - y_j}{n_j q_j}\right)\right]} \qquad (35)$$

where n_j = total number of subjects with the same x_j, y_j = number of positive responses (i.e., number times that y =1), and the sign of $d(y_j, p_j)$ is positive iff $(y_j - n_j p_j) > 0$. For the Challenger data at x = 70°F, $n_9 = 4$, $y_9 = 3$, $\hat{p}_9 = 0.24655$ and $\hat{q}_9 = 0.75345$. Hence, $d(y_9, 0.24655) = \sqrt{2\left[3\ln\left(\frac{3}{0.98621}\right) + (4-3)\ln\left(\frac{4-3}{3.013791}\right)\right]} = 2.11390607$, which is the same as that of Minitab's to 7 decimals. Note that when $n_j = 1$ and $y_j = 1$, then Eq. (35) reduces to $d[1, \hat{p}(x)]$, and when $y_j = 0$, Eq. (35) reduces to $d[0, \hat{q}(x)]$ and always $\sum_{\text{all x's}} n_j = n$; further, unlike residuals for most statistical models, the deviance residuals do not usually sum identically to zero,

i.e., their sum is generally different from zero but is usually close to zero. For the Challenger data, $\sum_{\text{all x's}} d(y_j, p_j)$ = – 3.41409357. The deviance in logistic regression is defined as D = $\sum_{\text{all Distinct x's}} [d(y_j, p_j)]^2$ = SS_{RES} (36)

and for the Challenger data D =15.75918096. The statistic D in Eq. (36) is asymptotically chi-square with df = number of distinct x_j's – 2 constraints. For the Challenger data there are 17 distinct x values, and hence, the *P-value* = $\hat{\alpha}$ = $\Pr(\chi^2_{15} \geq 15.75918096)$ = 0.39823473, which again implies that the logistic model fits the data well. Note that there are 17 squared terms in Eq. (36) and that is why there must be only one residual when there are replications at an x_j so that the df would be 17 – 2. Further, the larger the *P*-value is, the better is the fit.

The Pearson Chi-Square Statistic

The residual for this statistic is defined as ($y_j - \hat{y}_j$), where again there are two possibilities: (1) x_j is not repeated (i.e., only a single value of x_j), (2) At least two or more subjects with the same value of x_j. When there is only one subject at x_j, then the rv y_j has a Bernoulli distribution with success Pr, $p_j = p(x_j)$, $E(x_j) = p_j$ and variance p_jq_j. Hence, the residual $e_j = y_j - \hat{y}_j = y_j - \hat{p}_j$.

When an x_j ($n_j > 1$) is replicated, say n_j times, then at that x_j, the rv y_j has a binomial pmf with success Pr, p_j, $E(y_j) = n_j \times p_j$, $V(y_j) = n_j \times p_j \times q_j$ so that $e_j = y_j - \hat{y}_j = y_j - n_j \times \hat{p}_j$. For the Challenger data, there are n_9 = 4 replicates at x = 70°F but y_j = 3 successes, and hence, $e_9 = 3 - 4 \times 0.24655 = 2.01379114$ and the corresponding studentized residual is given by $r_9 = 2.01379114/\sqrt{4 \times 0.24655 \times 0.75345}$ = 2.33616437. Further, like in the case of deviance residuals, the sum of Pearson's residuals does not add to zero but is close to zero. My calculations for Table 8 of

my website show that the Pearson's GOF statistic is $\chi^2_{15} \cong \sum_{j=1}^{17} \frac{(y_j - n_j\hat{p}_j)^2}{n_j \times \hat{p}_j \times \hat{q}_j}$ = 14.0485605 $\rightarrow$ *P-value* = *p* = Pr($\chi^2_{15} \geq$ 14.0485605) = 0.52184944. This matches the Pr level reported by Minitab to 3 decimals. It seems that, in general, the Deviance provides a more powerful test (i.e., smaller *P-value*) that the Pearson's GOF statistic.

Exercise 8. The ICU (Intensive Care Unit) data from Hosmer & Lemeshow (2000) has many features versus a patience's age, x. The primary objective was to use logistic regression to predict survival Pr at the time of hospital discharge. One important dichotomous rv was Vital Status (STA) where 0 denoted Lived and 1 indicated that the patient died. The STA data for n = 200 patients are provided on my website. (a) Obtain the ML estimates of β_0 & β_1 and give the estimate of the logit h(x). Compute the odds for a random patient of age 60 and interpret its value. (b) Obtain the 95% CI's for β_0 , β_1 & the odds ratio. (c) Obtain the 95% CI for the logit at x = 60 and for p(60 years old) and interpret. (d) Test H_0: $\beta_1 = 0$ using the Z-statistic, LRS, and the ST by computing their *P*-values. (e) Use Excel to compute the GOF statistics D and that of Pearson's, computing their *P*-values.