11-9.1 Logistic Regression

Linear regression often works very well when the response variable is **quantitative.** We now consider the situation where the response variable takes on only two possible values, 0 and 1. These could be arbitrary assignments resulting from observing a **qualitative** response. For example, the response could be the outcome of a functional electrical test on a semiconductor device for which the results are either a "success," which means the device works properly, or a "failure," which could be due to a short, an open, or some other functional problem.

Suppose that the model has the form

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{11-51}$$

and the response variable Y_i takes on the values either 0 or 1. We will assume that the response variable Y_i is a **Bernoulli random variable** with probability distribution as follows:

Y_i	Probability
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Now since $E(\epsilon_i) = 0$, the expected value of the response variable is

$$\begin{aligned} E(Y_i) &= 1(\pi_i) + 0(1 - \pi_i) \\ &= \pi_i \end{aligned}$$

This implies that

$$E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i$$

This means that the expected response given by the response function $E(Y_i) = \beta_0 + \beta_1 x_i$ is just the probability that the response variable takes on the value 1.

There are some substantive problems with the regression model in Equation 11-51. First, note that if the response is binary, the error terms ϵ_i can only take on two values, namely,

$$\begin{aligned} \epsilon_i &= 1 - (\beta_0 + \beta_1 x_i) \qquad \text{when } Y_i = 1 \\ \epsilon_i &= -(\beta_0 + \beta_1 x_i) \qquad \text{when } Y_i = 0 \end{aligned}$$

Consequently, the errors in this model cannot possibly be normal. Second, the error variance is not constant, since

$$\begin{aligned} \sigma^2_{y_i} &= E\{Y_i - E(Y_i)\}^2 \\ &= (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i) \\ &= \pi_i (1 - \pi_i) \end{aligned}$$

Notice that this last expression is just

$$\sigma^2_{y_i} = E(Y_i)[1 - E(Y_i)]$$

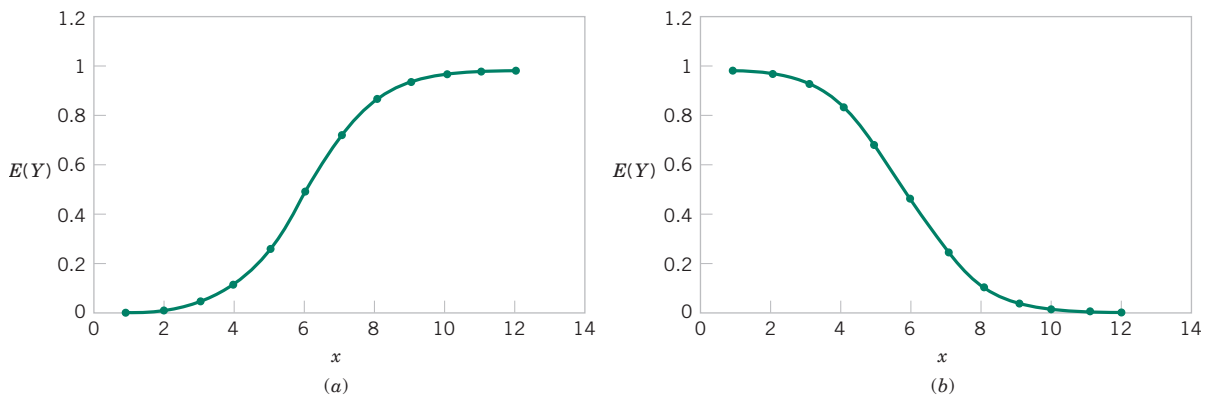

Figure 11-19 Examples of the logistic response function. (a) $E(Y) = 1/(1 + e^{-6.0-1.0x})$, (b) $E(Y) = 1/(1 + e^{-6.0+1.0x})$.

since $E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i$. This indicates that the variance of the observations (which is the same as the variance of the errors because $\epsilon_i = Y_i - \pi_i$, and π_i is a constant) is a function of the mean. Finally, there is a constraint on the response function, because

$$0 \le E(Y_i) = \pi_i \le 1$$

This restriction can cause serious problems with the choice of a **linear response function,** as we have initially assumed in Equation 11-51. It would be possible to fit a model to the data for which the predicted values of the response lie outside the 0, 1 interval.

Generally, when the response variable is binary, there is considerable empirical evidence indicating that the shape of the response function should be nonlinear. A monotonically increasing (or decreasing) *S*-shaped (or reverse *S*-shaped) function, such as shown in Figure 11-19, is usually employed. This function is called the **logit response function,** and has the form

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{11-52}$$

or equivalently,

$$E(Y) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x)]} \tag{11-53}$$

In **logistic regression** we assume that $E(Y)$ is related to x by the logit function. It is easy to show that

$$\frac{E(Y)}{1 - E(Y)} = \exp^{\beta_0 + \beta_1 x} \tag{11-54}$$

The quantity $\exp(\beta_0 + \beta_1 x)$ on the right-hand side of Equation 11-54 is called the **odds ratio.** It has a straightforward interpretation: If the odds ratio is 2 for a particular value of x, it means that a success is twice as likely as a failure at that value of the regressor x. Notice that the

natural logarithm of the odds ratio is a linear function of the regressor variable. Therefore the slope β_1 is the change in the log odds that results from a one-unit increase in x. This means that the odds ratio changes by e^{β_1} when x increases by one unit.

The parameters in this logistic regression model are usually estimated by the method of maximum likelihood. For details of the procedure, see Montgomery, Peck, and Vining (2001). Minitab will fit logistic regression models and provide useful information on the quality of the fit.

We will illustrate logistic regression using the data on launch temperature and O-ring failure for the 24 space shuttle launches prior to the *Challenger* disaster of January 1986. There are six O-rings used on the rocket motor assembly to seal field joints. The table below presents the launch temperatures. A 1 in the "O-Ring Failure" column indicates that at least one O-ring failure had occurred on that launch.

Temperature	O-Ring Failure	Temperature	O-Ring Failure	Temperature	O-Ring Failure
53	1	68	0	75	0
56	1	69	0	75	1
57	1	70	0	76	0
63	0	70	1	76	0
66	0	70	1	78	0
67	0	70	1	79	0
67	0	72	0	80	0
67	0	73	0	81	0

Figure 11-20 is a scatter plot of the data. Note that failures tend to occur at lower temperatures. The logistic regression model fit to this data from Minitab is shown in the following boxed display.

Binary Logistic Regression: O-Ring Failure versus Temperature

Link Function: Logit

Response Information

Variable	Value	Count	
O-Ring F	1	7	(Event)
	0	17	
	Total	24	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	10.875	5.703	1.91	0.057			
Temperat	−0.17132	0.08344	−2.05	0.040	0.84	0.72	0.99

Log-Likelihood = −11.515

Test that all slopes are zero: G = 5.944, DF = 1, P-Value = 0.015

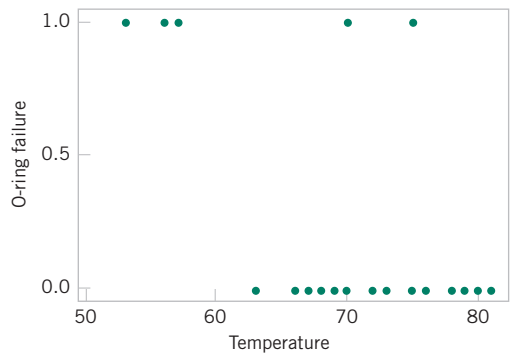

Figure 11-20 Scatter plot of O-ring failures versus launch temperature for 24 space shuttle flights.

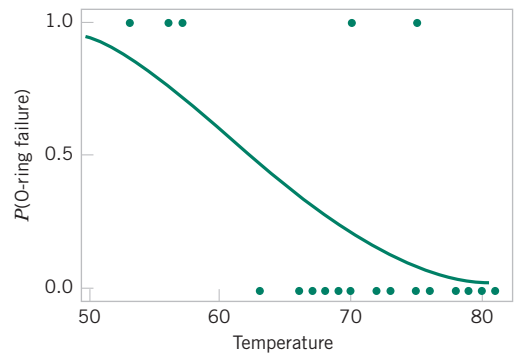

Figure 11-21 Probability of O-ring failure versus launch temperature (based on a logistic regression model).

The fitted logistic regression model is

$$\hat{y} = \frac{1}{1 + \exp[-(10.875 - 0.17132x)]}$$

The standard error of the slope $\hat{\beta}_1$ is $se(\hat{\beta}_1) = 0.08344$. For large samples, $\hat{\beta}_1$ has an approximate normal distribution, and so $\hat{\beta}_1/se(\hat{\beta}_1)$ can be compared to the standard normal distribution to test H_0: $\beta_1 = 0$. Minitab performs this test. The P-value is 0.04, indicating that temperature has a significant effect on the probability of O-ring failure. The odds ratio is 0.84, so every one degree increase in temperature reduces the odds of failure by 0.84. Figure 11-21 shows the fitted logistic regression model. The sharp increase in the probability of O-ring failure is very evident in this graph. The actual temperature at the *Challenger* launch was 31°F. This is well outside the range of other launch temperatures, so our logistic regression model is not likely to provide highly accurate predictions at that temperature, but it is clear that a launch at 31°F is almost certainly going to result in O-ring failure.

It is interesting to note that all of these data were available **prior** to launch. However, engineers were unable to effectively analyze the data and use them to provide a convincing argument against launching *Challenger* to NASA managers. Yet a simple regression analysis of the data would have provided a strong quantitative basis for this argument. This is one of the more dramatic instances that points out **why engineers and scientists need a strong background in basic statistical techniques.**