

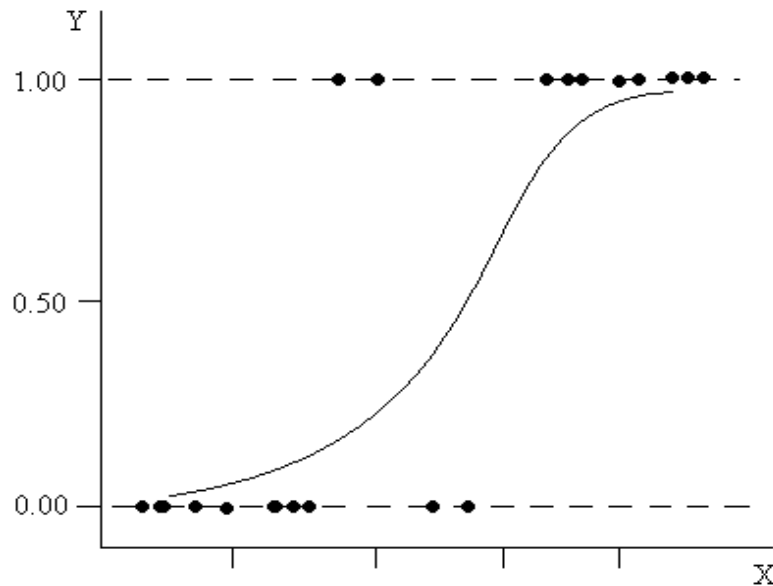
# Logistic Regression

Sara Vyrostek  
Senior Exercise  
November 16, 2001

## Introduction:

In the modeling of data, analysts develop relationships based upon the observed values of a set of predictor variables in order to determine the expected value of the response variable of interest, a technique known as regression. A very common form of regression analysis is linear regression, which examines the effect of a unit change in a predictor variable on the response variable. However, there are many cases, whether it is due to patterns in the data or the design of the data itself, where a linear analysis is inappropriate. One particular case where the linear analysis does not provide an adequate fit occurs when the response variable is binary, meaning it has only two possible values. For this scenario, a related technique known as logistic regression is used to model the data. The logistic model provides a curvilinear fit with asymptotes at both zero and one, the two possible values for the response variable.<sup>1</sup> Figure 1 displays a graph of a logistic model fit to a typical binary data set, visually displaying these characteristics.

Figure 1: A Typical Logistic Model



<sup>1</sup> Neter, John, William Wasserman, Christopher J. Nachtsheim, and Michael H. Kutner, *Applied Linear Regression Models*, 3<sup>rd</sup> ed. (Chicago: McGraw-Hill, 1996) p. 570.

The logistic regression technique is widely used in any area in which the response variable tends to take on the binary form. For example, in health research it is common for the response variable to have the form “yes or no” or “responded to medication versus did not respond to medication”, situations in which logistic regression would be employed. Not only will this paper serve as an introduction to regression in the logistic setting, but it will also step through the processes of parameter estimation and interpretation, the building of a logistic regression model, and the analysis of fit. Throughout the paper, an ongoing analysis of a data set by means of both hand calculations as well as the interpretation of SAS output will provide the opportunity to work through all of the topics discussed. The data set, a subset of variables and observations taken from a larger study done at the University of Massachusetts Aids Research Unit in order to examine drug use in association with HIV, was found in Hosmer and Lemeshow’s *Applied Logistic Regression* (2000). The variables in the data set are as follows:

**Figure 2: Description of Variables<sup>2</sup>**

| Variable | Description                        | Codes/Values                      | Name     |
|----------|------------------------------------|-----------------------------------|----------|
| 1        | Identification Code                | 1-575                             | ID       |
| 2        | Age at Enrollment                  | Years                             | age      |
| 3        | Beck Depression Score at Admission | 0.000-54.000                      | beck     |
| 4        | IV Drug Use History at Admission   | 1=Never, 2=Previous, 3=Recent     | IV       |
| 5        | Number of Prior Drug Treatments    | 0-40                              | priors   |
| 6        | Subject’s Race                     | 0=White, 1=Other                  | race     |
| 7        | Treatment Randomization Assignment | 0=Short, 1=Long                   | treatmnt |
| 8        | Treatment Site                     | 0=A, 1=B                          | site     |
| 9        | Remained Drug Free for 12 Months   | 1=Remained Drug Free, 0=Otherwise | drugFree |

<sup>2</sup> Hosmer, David W., and Stanley Lemeshow, *Applied Logistic Regression* (New York: John Wiley & Sons, 2000) pp. 26-27.

### Simple Logistic Regression Model:

Working in the binary setting,  $Y$  can take on only one of two values, and thus can be coded as follows:

$$Y_i = \begin{cases} 1, & \text{if the event occurs} \\ 0, & \text{if the event does not occur} \end{cases}$$

In the simple setting with only one predictor variable, the logistic regression model will take on the following general form:

$$Y_i = \pi(x_i) + \varepsilon_i$$

where,

$$\pi(x_i) = E[Y | X] = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}},$$

and  $\varepsilon_i$  is a random error term.<sup>3</sup> Alternatively,  $\pi(x_i)$  can be written as

$$\pi(x_i) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_i}}$$

Because  $\pi(x_i)$  represents the probability that  $Y_i = 1$  given the value of  $X$ , this implies that  $1 - \pi(x_i)$  represents the probability that  $Y_i = 0$ , given the value of  $X$ .  $Y_i$  is distributed such that it has two possible outcomes with a probability of success  $\pi(x_i)$ , meaning that the  $Y_i$ 's follow the Bernoulli distribution. Unlike the linear model where the error terms are normally distributed, the error terms in this model have only two possible values. If  $Y_i = 1$ , then  $\varepsilon_i = 1 - \pi(x_i)$  with probability  $\pi(x_i)$ , corresponding to the above information. On the other hand, if  $Y_i = 0$  which occurs with probability  $1 - \pi(x_i)$ , then  $\varepsilon_i = -\pi(x_i)$ . The mean of the error terms can be found by calculating their expected value. Since  $\varepsilon_i$  can only take on two possible values, the expected value is obtained through the following formula:

---

<sup>3</sup> Chatterjee, Samprit, Ali S. Hadi, and Bertram Price, *Regression Analysis by Example* (New York: John Wiley & Sons, 2000) p. 320.

$$\begin{aligned}
E(\varepsilon_i) &= \varepsilon_1 P(\varepsilon_i = \varepsilon_1) + \varepsilon_2 P(\varepsilon_i = \varepsilon_2) \\
&= [1 - \pi(x_i)]\pi(x_i) + [-\pi(x_i)][1 - \pi(x_i)] \\
&= \pi(x_i) - \pi(x_i)^2 - \pi(x_i) + \pi(x_i)^2 \\
&= 0
\end{aligned}$$

The variance of the error terms can be determined in a similar manner:

$$\begin{aligned}
Var(\varepsilon_i) &= E(\varepsilon_i^2) - [E(\varepsilon_i)]^2 \\
&= [\varepsilon_1^2 P(\varepsilon_i = \varepsilon_1) + \varepsilon_2^2 P(\varepsilon_i = \varepsilon_2)] - 0 \\
&= [1 - \pi(x_i)]^2 \pi(x_i) + [-\pi(x_i)]^2 [1 - \pi(x_i)] \\
&= \pi(x_i)[1 - \pi(x_i)] \{ [1 - \pi(x_i)] + \pi(x_i) \} \\
&= \pi(x_i)[1 - \pi(x_i)]
\end{aligned}$$

This information provides a basic understanding of the distributional setting of this analysis.<sup>4</sup>

Having defined and described the logistic regression function, the next task involves estimating the parameters  $\beta_0$  and  $\beta_1$  for the model when fitting it to an actual data set. By the very nature of the probability function described above, the estimation process can be simplified by forming a linear equation involving these two parameters. With  $\pi(x_i)$  denoting the probability that  $Y_i = 1$ , the odds of Y occurring can be defined as the probability that  $Y_i = 1$  divided by the probability that  $Y_i \neq 1$ , or  $Y_i = 0$ . In other words,

$$\text{odds} = \frac{\pi(x_i)}{1 - \pi(x_i)}.$$

The odds are important for a binary data set because they offer a ratio for the chances of an event occurring as opposed to an event not occurring. For example, if the probability of the sun coming out on a given day is two-thirds, meaning that the probability of not seeing the sun on a given day is one-third, then the odds of a sunny day are  $\frac{(2/3)}{(1/3)}$ , or a two to one ratio. This

---

<sup>4</sup> Chatterjee, p. 320.

implies that a sunny day is twice as likely as a cloudy day. Because the equation for the odds results in a difficult function to fit,  $h(x)$  can be defined as the natural log of the odds, also known as the logit function:

$$\begin{aligned}
 h(x_i) &= \ln \left[ \frac{\pi(x_i)}{1 - \pi(x_i)} \right] \\
 &= \ln \left[ \frac{(1 + e^{-\beta_0 - \beta_1 x_i})^{-1}}{1 - (1 + e^{-\beta_0 - \beta_1 x_i})^{-1}} \right] \\
 &= \ln \left[ \frac{1}{e^{-\beta_0 - \beta_1 x_i}} \right] \\
 &= \ln(e^{\beta_0 + \beta_1 x_i}) \\
 &= \beta_0 + \beta_1 x_i
 \end{aligned}$$

This function,  $h(x_i)$ , is of the linear form, providing a much easier model to fit.<sup>5</sup>

The parameters  $\beta_0$  and  $\beta_1$  can be estimated through the method of maximum likelihood estimation. This method isolates the values of a parameter that maximize the likelihood function, where the likelihood function  $L(\beta_0, \beta_1)$  is defined as the joint probability distribution for all of the data points.<sup>6</sup> Since the  $Y_i$ 's have a Bernoulli distribution, the probability density function can be defined as

$$P(Y = y_i) = f_i(y_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1 - y_i}, \text{ where } y_i = 0 \text{ or } 1, \text{ and } i = 1, \dots, n.$$

Because the  $Y_i$ 's are independent, the likelihood function can be defined as follows:

$$\begin{aligned}
 L(\beta_0, \beta_1) &= g(y_1, \dots, y_n) \\
 &= \prod_{i=1}^n f_i(y_i) \\
 &= \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1 - y_i}
 \end{aligned}$$

---

<sup>5</sup> Hosmer, David W., and Stanley Lemeshow, *Applied Logistic Regression* (New York: John Wiley & Sons, 1989) p. 6.

<sup>6</sup> Hosmer, p. 8.

In order to maximize this function, the derivative must be taken with respect to each of the parameters. Then, the resulting equations would be set equal to zero and solved simultaneously. This process can be simplified by performing the same analysis on the natural log of the likelihood function, being that maximizing the natural log of the function would result in the same value as maximizing the likelihood function itself. Obtaining the log-likelihood function:

$$\begin{aligned}
\ln L(\beta_0, \beta_1) &= \ln \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \\
&= \sum_{i=1}^n \left\{ \ln \left[ \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \right] \right\} \\
&= \sum_{i=1}^n \left\{ \ln \left[ \pi(x_i)^{y_i} \right] + \ln \left[ (1 - \pi(x_i))^{1-y_i} \right] \right\} \\
&= \sum_{i=1}^n \left\{ y_i \ln \left[ \pi(x_i) \right] + (1 - y_i) \ln \left[ 1 - \pi(x_i) \right] \right\} \\
&= \sum_{i=1}^n \left\{ y_i \ln \left( \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) + (1 - y_i) \ln \left[ 1 - \left( \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \right] \right\} \\
&= \sum_{i=1}^n \left\{ y_i \left[ \beta_0 + \beta_1 - \ln(1 + e^{\beta_0 + \beta_1 x_i}) \right] + (1 - y_i) \left[ -\ln(1 + e^{\beta_0 + \beta_1 x_i}) \right] \right\} \\
&= \sum_{i=1}^n \left\{ y_i (\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i}) \right\}
\end{aligned}$$

Now taking the derivative, first with respect to  $\beta_0$  and then with respect to  $\beta_1$  and setting each equal to zero, the following likelihood equations are formed:

$$\begin{aligned}
\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_0} &= \sum_{i=1}^n \left\{ y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right\} = 0 & \frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_1} &= \sum_{i=1}^n \left\{ y_i x_i - \frac{x_i e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right\} = 0 \\
&= \sum_{i=1}^n [y_i - \pi(x_i)] = 0 & &= \sum_{i=1}^n x_i [y_i - \pi(x_i)] = 0
\end{aligned}$$

Because the likelihood equations are not linear, solving these equations simultaneously requires an iterative procedure that is normally left to a software package.<sup>7</sup>

---

<sup>7</sup> Neter, pp. 574-575.

Let us consider the data set on drug use. Suppose a simple model is desired using only one predictor variable,  $t$ , where  $t$  represents the type of treatment administered. We will designate the response variable  $Y$  as an indicator for whether or not the subject remained drug free for twelve months. The basic logistic model is:

$$Y_i = \pi(t_i) + \varepsilon_i = \frac{e^{\beta_0 + \beta_1 t_i}}{1 + e^{\beta_0 + \beta_1 t_i}} + \varepsilon_i.$$

The accompanying logit function is:

$$h(t_i) = \ln \left[ \frac{\pi(t_i)}{1 - \pi(t_i)} \right] = \beta_0 + \beta_1 t_i.$$

Using SAS to obtain the appropriate parameter estimates, the following output is produced:

**Display 1: Simple Logistic Regression--Parameter Estimates**  
The LOGISTIC Procedure  
Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------|------------|
| Intercept | 1  | -1.2978  | 0.1433         | 82.0211         | <.0001     |
| treatmnt  | 1  | 0.4371   | 0.1931         | 5.1266          | 0.0236     |

| Effect   | Odds Ratio Estimates |                            |  |
|----------|----------------------|----------------------------|--|
|          | Point Estimate       | 95% Wald Confidence Limits |  |
| treatmnt | 1.548                | 1.060 2.260                |  |

Display 1 reports the value for  $b_0$ , the parameter estimate for  $\beta_0$ , as  $-1.2978$ , and the value for  $b_1$ , the parameter estimate for  $\beta_1$ , as  $0.4371$ . Thus, the estimated logit function would read

$$h(t_i) = -1.2978 + 0.4371t_i,$$

with the parameter estimates for the model having been determined through the method of maximum likelihood estimation. The value for the log-likelihood function is also produced in the SAS output:

**Display 2: Simple Logistic Regression--The Log-Likelihood Function**  
Model Fit Statistics

Intercept      Intercept  
and

| Criterion | Only    | Covariates |
|-----------|---------|------------|
| AIC       | 655.729 | 652.551    |
| SC        | 660.083 | 661.259    |
| -2 Log L  | 653.729 | 648.551    |

By interpreting the output in Display 2, it is evident that the log-likelihood function is equal to  $648.551/-2 = -324.2755$ . The output also reports the value of the log-likelihood function for the model with just the intercept, which is equal to  $653.729/-2 = -326.8645$ .

Since the logit function is written in a linear form with  $\beta_1$  representing the slope,  $b_1$  represents the change in  $h(x)$  for a one-unit change in  $x$ . Therefore,

$$\begin{aligned}
 b_1 &= h(x+1) - h(x) \\
 &= \ln\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \\
 &= \ln(odds_2) - \ln(odds_1) \\
 &= \ln\left(\frac{odds_2}{odds_1}\right).
 \end{aligned}$$

Taking the exponential of both sides, we get:

$$e^{b_1} = \frac{odds_2}{odds_1},$$

which can be defined as the odds ratio. Therefore, after finding the parameter estimates, the value  $e^{b_1}$  will represent the percentage increase in the probability that  $Y = 1$  for each unit increase in  $X$ .<sup>8</sup> This value for the drug data is seen in Display 1, with a point estimate for the odds ratio of 1.548. This indicates that by changing the treatment period from short to long, the probability of a subject remaining drug free for a twelve-month period after the treatment increases by 54.8%.

Despite having fit the model and interpreted the coefficients, this information is useless unless the predictor variable is significant to the model. One method for determining whether or

---

<sup>8</sup> Neter, p. 577.

not the coefficient is significant is through the likelihood ratio test. Assuming that we have  $n$  observations on each variable, the deviance of the model,  $D$ , will be defined as follows:

$$D = -2 \ln \left[ \frac{\text{likelihood of the current Model}}{\text{likelihood of the saturated Model}} \right].$$

The saturated model represents the model containing  $n$  parameters, such that the model perfectly predicts the observed data set. In other words, the predicted values for this model are equal to the observed values in the data set. The deviance provides a means of comparing the likelihood of the model that has been fit, or the probability of obtaining the observed data set given the current model, to that of the saturated model.<sup>9</sup> By manipulating the above definition, we find:

$$D = -2[\ln(\text{likelihood of the current Model}) - \ln(\text{likelihood of the saturated Model})].$$

Recall that the general model is of the form  $Y_i = \pi(x_i) + \varepsilon_i$ . For the current model,  $\hat{\pi}(x_i)$  serves as the estimator for  $\pi(x_i)$ . However, the saturated model generates the complete data set such that  $y_i$  serves as the estimator for  $\pi(x_i)$ . Also recalling the derivation of the likelihood function, we know that:

$$\begin{aligned} \ln L(\beta_0, \beta_1) &= \ln \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \\ &= \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \end{aligned}$$

This equation can be substituted into the formula for the deviance and then manipulated in order to get the following working equation:

---

<sup>9</sup> Hosmer, p. 13.

$$\begin{aligned}
D &= -2 \left\{ \left[ \sum_{i=1}^n (y_i \ln[\hat{\pi}(x_i)] + (1 - y_i) \ln[1 - \hat{\pi}(x_i)]) \right] - \left[ \sum_{i=1}^n (y_i \ln(y_i) + (1 - y_i) \ln(1 - y_i)) \right] \right\} \\
&= -2 \left\{ \left[ \sum_{i=1}^n (y_i \ln[\hat{\pi}(x_i)] - y_i \ln(y_i) + (1 - y_i) \ln[1 - \hat{\pi}(x_i)] - (1 - y_i) \ln(1 - y_i)) \right] \right\} \\
&= -2 \sum_{i=1}^n \left[ y_i \ln \left( \frac{\hat{\pi}(x_i)}{y_i} \right) + (1 - y_i) \ln \left( \frac{1 - \hat{\pi}(x_i)}{1 - y_i} \right) \right].^{10}
\end{aligned}$$

Essentially, the deviance takes the likelihood of the current model where an element of error is present and subtracts the likelihood of the saturated model in which there is no error term present, and then sums over this difference. Thus, the deviance in the logistic setting serves the same purpose as the residual sums of squares in the linear setting.

In order to determine whether or not the parameter is significant to the model, the deviance of the model containing the parameter must be compared with the deviance of the model without the parameter. Therefore, the test statistic,  $G$ , is:

$$G = D(\text{for the model without the variable}) - D(\text{for the model with the variable})$$

$$G = -2 \ln \left( \frac{\text{likelihood of current Model without } \beta_1}{\text{likelihood of saturated Model}} \right) + 2 \ln \left( \frac{\text{likelihood of current Model with } \beta_1}{\text{likelihood of saturated Model}} \right)$$

$$G = -2 \ln \left( \frac{\text{likelihood of current Model without } \beta_1}{\text{likelihood of current Model with } \beta_1} \right)$$

$$G = -2 \ln(\text{likelihood of current Model without } \beta_1) + 2 \ln(\text{likelihood of current Model with } \beta_1)^{11}$$

We have already found the likelihood function for the model with the single predictor variable. For the likelihood function for the model without the single predictor variable, the probability that  $Y_i = 0$  will be equal to the average number of zeros in the sample, while the probability that  $Y_i = 1$  will be equal to the average number of ones in the sample. Therefore,

---

<sup>10</sup> Hosmer, p. 13.

<sup>11</sup> Hosmer, pp. 13-14.

$$\begin{aligned}
L(\beta_0) &= \bar{y}^{\sum_{i=1}^n y_i} (1 - \bar{y})^{\sum_{i=1}^n (1 - y_i)} \\
&= \left( \frac{\sum_{i=1}^n y_i}{n} \right)^{\sum_{i=1}^n y_i} \left( \frac{\sum_{i=1}^n (1 - y_i)}{n} \right)^{\sum_{i=1}^n (1 - y_i)}
\end{aligned}$$

Given that likelihood function, we can determine the appropriate equation for the test statistic G:

$$G = -2 \ln \left[ \frac{\left( \frac{\sum_{i=1}^n y_i}{n} \right)^{\sum_{i=1}^n y_i} \left( \frac{\sum_{i=1}^n (1 - y_i)}{n} \right)^{\sum_{i=1}^n (1 - y_i)}}{\prod_{i=1}^n \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{(1 - y_i)}} \right].$$

G follows the chi-squared distribution with one degree of freedom. In checking the significance of the coefficient, the following null and alternative hypotheses are to be tested:

$$\begin{aligned}
H_0 &: \beta_1 = 0 \\
H_a &: \beta_1 \neq 0.
\end{aligned}$$

For this test, the decision rule requires that if  $G > \chi^2(1 - \alpha; 1)$ ,  $H_0$  is to be rejected, meaning that the coefficient would be deemed significant. On the other hand if  $G \leq \chi^2(1 - \alpha; 1)$ , then one must fail to reject  $H_0$ , concluding that the coefficient is insignificant.<sup>12</sup> SAS reports the test statistic and the corresponding p-value, where the p-value is equal to the probability of observing a test statistic at least as extreme as the observed value assuming that the null hypothesis is true. For this likelihood ratio test, the p-value =  $P(G \geq \text{the observed G-value})$ , where  $G \sim \chi^2(1 - \alpha; 1)$ .

Because the  $\alpha$ -level represents the probability of rejecting the null hypothesis when the null hypothesis is true, as long as the p-value is less than the chosen  $\alpha$ -level, the null hypothesis can be safely rejected. For all of the future tests in this paper, we will use an  $\alpha$ -level of 0.05.

For the drug data set, G can be calculated from the SAS output in two manners. First, knowing the formula for G, we see that:

$$G = -2\ln(\text{likelihood of current Model without } \beta_1) + 2\ln(\text{likelihood of current Model with } \beta_1).$$

The values for  $-2*\log$ -likelihood entries from Display 2 can be inserted to calculate the test statistic. Using this information,  $G = 653.729 - 648.551 = 5.178$ . Therefore with an alpha level of 0.05 and  $\chi^2(1 - 0.05; 1) = 3.8415$ , the null hypothesis can be rejected, meaning that the parameter estimate for the type of treatment administered is significant to the model. The p-value for this test represents the probability of observing a value for G of at least 5.178 such that the p-value =  $P(G \geq 5.178) = 0.0229$ . Because this p-value is less than the desired alpha level of 0.05, the null hypothesis would be rejected here, leading to the same conclusion. Another section of the SAS output actually computes G, or the likelihood ratio test statistic, and calculates the appropriate p-value for the test of the above hypotheses:

**Display 3: Simple Logistic Regression--The Likelihood Ratio Test**  
Testing Global Null Hypothesis: BETA=0

| Test             | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 5.1782     | 1  | 0.0229     |
| Score            | 5.1626     | 1  | 0.0231     |
| Wald             | 5.1266     | 1  | 0.0236     |

The row reading likelihood ratio reports the test statistic  $G = 5.1782$ , the same as that calculated above, with one degree of freedom, and a p-value of 0.0229.

---

<sup>12</sup> Hosmer, pp. 12-14.

Another method for computing the significance of a coefficient is through the Wald test, where in order to test the same hypotheses as above, the test statistic  $W = \frac{b_k}{s(b_k)}$  is used. Since this test statistic is approximately normal, if  $|W| \leq z(1 - \alpha/2)$  then one must fail to reject the null hypothesis, while if  $|W| > z(1 - \alpha/2)$ , then the null hypothesis can be rejected at the given alpha level.<sup>13</sup> For this test, the p-value can be defined as follows:

$$\begin{aligned} p - value &= P(|W| > \text{the observed test statistic}) \\ &= 2P(W > \text{the observed test statistic}) \end{aligned}$$

where  $W \sim z(1 - \alpha/2)$ . For the drug data set,  $W = \frac{0.4371}{0.1931} = 2.2636$ . Using an alpha level of

0.05, the critical value is  $z(0.975) = 1.96$ , and the corresponding p-value is

$P(|W| > 2.2636) = 2 * .0118 = 0.0236$ . Since  $|W| > z$  and the corresponding p-value is less than 0.05, the null hypothesis will still be rejected and same conclusion of variable significance can be reached.

This test also can be written in an alternative manner. Because the squaring a normal random variable will result in a chi-square random variable with one degree of freedom<sup>14</sup>, the

Wald test statistic can be written as  $W^2 = \left(\frac{b_k}{s(b_k)}\right)^2$  where  $W^2 \sim \chi^2(1 - \alpha; 1)$ . In accordance

with this change, the decision rule must be adjusted such that the null hypothesis is rejected

when  $W^2 > \chi^2(1 - \alpha; 1)$ . Likewise, the p-value will be redefined so that the

$p - value = P(W^2 > \text{the observed test statistic})$ . Looking back at Display 1, we can see that SAS reports this test statistic, rather than those previously described. For the single predictor

---

<sup>13</sup> Neter, pp. 601-602.

variable *treatment*, the Wald chi-square test statistic is 5.1266 with a corresponding p-value of 0.0236, again indicating that this predictor variable is significant to the model.

**Multiple Logistic Regression Model:**

We can extend the above analysis for the simple logistic regression model where we have only one predictor variable to a setting with more than one predictor variable. In this setting, the vector  $\vec{x} = (x_1 + x_2 + \dots + x_p)$  represents the collection of  $p$  predictor variables for this model.

The formulas for the probability that  $Y_i = 1$ ,  $\pi(x_i)$ , as well as for the logit transformation,  $h(x_i)$ , can be extended where:

$$\pi(\vec{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \text{ and}$$

$$h(\vec{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p .$$

The only time that the model will differ from this general formula is if one of the predictor variables is a discrete, categorical variable with more than two levels. If one or more of these is present and if the number of variable categories is equal to  $k$ , then  $k-1$  design variables must be created.<sup>15</sup> These design variables are just binary variables meant to serve as an indicator for one of the levels of the associated variable. Using our drug data variable *IV* as an example, this variable would have to be recoded because it is a discrete variable with three levels: recent, previous, or no history of IV drug use. Because *IV* currently has three discrete levels, two indicator variables would have to be created and substituted into the model. These two new variables might be *IV1* and *IV2*, where

$$IV1 = \begin{cases} 1, & IV = 3 \\ 0, & IV \neq 3 \end{cases}, \text{ and } IV2 = \begin{cases} 1, & IV = 2 \\ 0, & IV \neq 2 \end{cases} .$$

---

<sup>14</sup> Hogg, Robert V., and Allen T. Craig, *Introduction to Mathematical Statistics* (London: The MacMillan Company, 1970) p. 109.

<sup>15</sup> Hosmer, pp. 25-27.

We would only be required to include two indicator variables because the third value of the predictor variable is implicitly included in the model. In this case, when  $IV1 = 1$  (which means that  $IV2 = 0$ ), this indicates that the subject has no history of drug use. Likewise, when  $IV2 = 1$  (which means that now  $IV1 = 0$ ), this means that the subject has a history of previous drug use. However, if both of these indicator variables equal zero, this would imply that the subject has a history of neither no drug use nor previous drug use. Instead, this would indicate that the subject has a history of recent drug use.

The parameters in the multiple setting are once again determined through maximum likelihood estimation. Because  $Y$  still remains a Bernoulli variable with the same probability distribution, the derivation of the maximum likelihood estimators remains the same, with the exception of the inclusion of more parameters. Thus, the log-likelihood equation would take the form:

$$\ln L(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n \left\{ y_i (\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}) - \ln(1 + e^{\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}}) \right\}.$$

In the same manner as before, the equations resulting from taking the derivative of the log-likelihood equation with respect to each of the parameters and then setting each derivative equal to zero would be solved simultaneously in order to obtain the estimates. Because this procedure is even more computationally intensive with multiple parameters, the estimation once again is left to computer software.<sup>16</sup>

Examining multiple logistic regression by means of the drug use data set, we can develop a model including three chosen predictor variables: *age*, *race*, and *treatment*. The model will take the form:

---

<sup>16</sup> Neter, p. 27.

$$Y_i = \pi(\bar{x}_i) + \varepsilon_i, \text{ where } \pi(\bar{x}_i) = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}}},$$

with  $x_1$  representing *age*,  $x_2$  *race*, and  $x_3$  *treatment*. The logit will take the form:

$$h(\bar{x}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}.$$

Using SAS to estimate the appropriate parameters, the following information is printed in the output:

**Display 4: Multiple Logistic Regression--Parameter Estimates**  
Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------|------------|
| Intercept | 1  | -2.0432  | 0.5350         | 14.5848         | 0.0001     |
| age       | 1  | 0.0196   | 0.0155         | 1.6016          | 0.2057     |
| race      | 1  | 0.4230   | 0.2126         | 3.9590          | 0.0466     |
| treatmnt  | 1  | 0.4218   | 0.1947         | 4.6942          | 0.0303     |

Odds Ratio Estimates

| Effect   | Point Estimate | 95% Wald Confidence Limits |       |
|----------|----------------|----------------------------|-------|
| age      | 1.020          | 0.989                      | 1.051 |
| race     | 1.527          | 1.006                      | 2.316 |
| treatmnt | 1.525          | 1.041                      | 2.233 |

Using these values, the estimated logit equation is:

$$h(\bar{x}_i) = -2.0432 + 0.0196x_{1i} + 0.4230x_{2i} + 0.4218x_{3i}.$$

The interpretation of the parameters is the same here as for the simple model. The odds ratio estimate, or  $e^{b_i}$ , represents the percentage increase or decrease in the probability that a drug user remains drug free 12 months after the conclusion of the treatment program due to a one-unit change in the variable  $x_i$ , holding all else constant. That percentage change in probability is equal to the odds ratio estimate minus one. According to the point estimates for the odds ratios given in the output above, the likelihood of a subject remaining drug free upon the conclusion of treatment increases by 2.0% for each additional year in age. Also, the odds that a non-white person remains drug free are 52.7% greater than the odds of a white person remaining drug free.

Finally, a person receiving a long treatment as opposed to a short treatment would be about 1.5 times more likely to remain drug free as well.

Again, before drawing any conclusions from the estimated logistic regression model, the complete model as well as the individual coefficients should be tested for significance. First, using the deviance previously described, the significance of the model can be tested. The hypotheses of interest are:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_A : \text{at least one of the above } \beta\text{'s} \neq 0$$

In order to test this, we want to compute G where

$$G = D(\text{for the model containing } \beta_0, \text{ or the reduced model}) - D(\text{for the full model})$$

$$= -2 \ln \left( \frac{\text{likelihood of the reduced Model}}{\text{likelihood of the saturated Model}} \right) + 2 \ln \left( \frac{\text{likelihood of the full Model}}{\text{likelihood of the saturated Model}} \right)$$

$$= -2 \ln \left( \frac{\text{likelihood of the reduced Model}}{\text{likelihood of the full Model}} \right)$$

$$= -2 \ln(\text{likelihood of the reduced Model}) + 2 \ln(\text{likelihood of the full Model}).^{17}$$

From the SAS output,

**Display 5: Multiple Logistic Regression—Significance of Model**  
Model Fit Statistics

| Criterion | Intercept Only | Intercept and Covariates |
|-----------|----------------|--------------------------|
| AIC       | 655.729        | 651.023                  |
| SC        | 660.083        | 668.441                  |
| -2 Log L  | <b>653.729</b> | <b>643.023</b>           |

Thus,

$$G = 653.729 - 643.023 = 10.706 .$$

Notice that the values in the “intercept only” column are identical to those values in Display 2.

Because this column is independent of the number or type of predictor variables put into the

---

<sup>17</sup> Hosmer, p. 31.

model, this column will remain the same for whatever set of predictor variables are included. The computed G-statistic is echoed in Display 6 below, another portion of the SAS output that includes the likelihood ratio test:

**Display 6: Multiple Logistic Regression--The Likelihood Ratio Test**  
Testing Global Null Hypothesis: BETA=0

| Test             | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 10.7055    | 3  | 0.0134     |
| Score            | 10.7845    | 3  | 0.0129     |
| Wald             | 10.5775    | 3  | 0.0142     |

G will once again follow the chi-square distribution, but in this case with three degrees of freedom because the significance of three parameters is being tested. The same decision rules apply to this setting as to the simple setting above. Therefore, the null hypothesis will be rejected if  $G > \chi^2(1 - \alpha; 3)$  and the null hypothesis will not be rejected if  $G \leq \chi^2(1 - \alpha; 3)$ . With  $\chi^2(0.95; 3) = 7.8147$  and the associated p-value = 0.0134, there is enough evidence to reject the null hypothesis and conclude that the parameters are significant.

Knowing that the entire model is significant does not guarantee the significance of all of the parameters included in the model. Therefore, the parameters should be tested individually or in small groups using either the likelihood-ratio test or the Wald test, already described, in order to check for their significance. In using the likelihood ratio to test the significance of a group of variables, the G statistic would be obtained by comparing the deviance of the full model, or that containing all of the parameters, to the deviance of the reduced model, or that with the parameters in question having been eliminated. The resulting test statistic would then be compared with the chi-square critical value based upon degrees of freedom equaling the number of parameters being tested. The same decision rules already described would be used. For example, in the three variable model fit above, the significance of the individual parameters can be tested using the Wald p-values given in the SAS output from Display 4. Only one parameter

*age* with a p-value of 0.2057 appears to be insignificant at the  $\alpha = 0.05$  level. Knowing this information, *age* could be eliminated from the set of predictor variables and the model could be refitted. In testing the significance of a discrete categorical variable divided into various sublevels, a clear decision can be made only when the same test result occurs for all levels of the variable. Otherwise, a large element of uncertainty accompanies any conclusion.<sup>18</sup>

### **Building the Regression Model:**

There are two main statistical methods that are often employed in determining which variables to include in a model, the stepwise regression method and the best subset method. However, in addition to the statistical development of the model, it is important to have a substantial amount of knowledge about the topic being studied. In some instances, a variable that might be considered very important in terms of practicality would be eliminated statistically and would have to be forced into the model. Therefore while these procedures provide a good tool for the development of a working model, it is a good idea to consider the entry of all possible variables based upon a working knowledge of the data.<sup>19</sup>

The first technique, stepwise regression, begins with a base model containing only the intercept parameter. It then adds variables significant to the model until there are no remaining significant variables left to be added. The nice feature of this procedure is that at each step after a variable has been added to the model, all of the variables included in previous steps are retested in order to see if they are still significant. The inclusion and extraction of variables from the model in the stepwise procedure is based upon the likelihood-ratio test. Normally when doing the likelihood-ratio test, a commonly accepted alpha level such as 0.01 or 0.05 is chosen as the critical value for the entry of variables into the model. For the model building process, this

---

<sup>18</sup> Hosmer, p. 27.

<sup>19</sup> Hosmer, p. 83.

cutoff value for the entry or removal of a variable should be increased to around 0.20. This will help in avoiding possibly significant variables from being overlooked or removed unnecessarily from the model.<sup>20</sup>

Using the stepwise procedure to build a model for our drug use data set, the process and results will be discussed with the aid of SAS output. The response variable for this model will be  $Y = drugfree$ , where  $Y$  is an indicator variable for whether or not a subject remains drug free for twelve months. The possible predictor variables will include the complete pool of variables available, where  $x_1 = age$ ,  $x_2 = beck$ ,  $x_3 = IV1$ ,  $x_4 = IV2$ ,  $x_5 = priors$ ,  $x_6 = race$ ,  $x_7 = treatmnt$ , and  $x_8 = site$ . Initially a model is fit with just the intercept term and then the corresponding likelihood,  $L(\beta_0)$ , is computed.

**Display 7: Step 1-- The intercept is entered.**

|   |            |          |            |            |            |
|---|------------|----------|------------|------------|------------|
| Model Convergence Status                      |            |          |            |            |            |
| Convergence criterion (GCONV=1E-8) satisfied. |            |          |            |            |            |
| Analysis of Maximum Likelihood Estimates      |            |          |            |            |            |
|   |            |          | Standard   | Wald       |            |
| Parameter                                     | DF         | Estimate | Error      | Chi-Square | Pr > ChiSq |
| Intercept                                     | 1          | -1.0687  | 0.0956     | 124.9675   | <.0001     |
| Residual Chi-Square Test                      |            |          |            |            |            |
|   | Chi-Square | DF       | Pr > ChiSq |            |            |
|   | 32.6798    | 8        | <.0001     |            |            |

SAS then fits eight individual regression models, each containing just one of the eight available predictor variables, and computes the likelihood  $L(\beta_0, \beta_i)$  for each of these models. Using these likelihood values, a Score chi-square test statistic is found by means of a fairly involved matrix computation. It involves a comparison of the full model versus the reduced model, where the full model contains the parameter(s) whose significance you are testing, while the reduced model does not include these parameters. The definition for the Score statistic follows for a full model containing  $s$  parameters and a reduced model containing  $t$  parameters:

---

<sup>20</sup> Hosmer, pp. 106-109.

$$Score = \bar{U}'(\hat{\beta}_0) \bar{I}^{-1}(\hat{\beta}_0) \bar{U}(\hat{\beta}_0)$$

where

$$\begin{aligned} \bar{\beta}_0 &= (\beta_0, \beta_1, \dots, \beta_t), \\ \bar{U}(\bar{\beta}_0) &= \left( \frac{\partial \ln L(\beta_0, \dots, \beta_t)}{\partial \beta_0}, \dots, \frac{\partial \ln L(\beta_0, \dots, \beta_t)}{\partial \beta_t} \right), \\ \bar{I}(\bar{\beta}_0) &= - \begin{bmatrix} \frac{\partial \left( \frac{\partial \ln L(\beta_0, \dots, \beta_t)}{\partial \beta_0} \right)}{\partial \beta_0} & \dots & \frac{\partial \left( \frac{\partial \ln L(\beta_0, \dots, \beta_t)}{\partial \beta_0} \right)}{\partial \beta_t} \\ \vdots & \ddots & \vdots \\ \frac{\partial \left( \frac{\partial \ln L(\beta_0, \dots, \beta_t)}{\partial \beta_t} \right)}{\partial \beta_0} & \dots & \frac{\partial \left( \frac{\partial \ln L(\beta_0, \dots, \beta_t)}{\partial \beta_t} \right)}{\partial \beta_t} \end{bmatrix} \end{aligned}$$

The Score statistic follows a chi-square distribution with  $(s - t)$  degrees of freedom.<sup>21</sup> This statistic for all of the possible variables, as well as the associated p-values, can be seen in

Display 8:

**Display 8: The Stepwise Procedure--Analysis of the Effects not in the Model**

| Effect        | DF       | Score<br>Chi-Square | Pr > Chisq    |
|---------------|----------|---------------------|---------------|
| age           | 1        | 1.4063              | 0.2357        |
| beck          | 1        | 0.6331              | 0.4262        |
| IV1           | 1        | 9.7368              | 0.0018        |
| IV2           | 1        | 0.2072              | 0.6490        |
| <b>priors</b> | <b>1</b> | <b>9.7585</b>       | <b>0.0018</b> |
| race          | 1        | 4.7791              | 0.0288        |
| treatmnt      | 1        | 5.1626              | 0.0231        |
| site          | 1        | 1.6921              | 0.1933        |

In general, a higher Score statistic is better than a lower one. With *priors* having the largest significant test statistic, this variable will be entered into the model. SAS outputs the parameter estimates as well as the likelihood values for the new model containing one predictor variable, *priors*. This new information appears in the following display.

**Display 9: Step 1--Priors is entered into the model**

Model Convergence Status  
Convergence criterion (GCONV=1E-8) satisfied.

| Criterion | Model Fit Statistics |                                |
|-----------|----------------------|--------------------------------|
|           | Intercept<br>Only    | Intercept<br>and<br>Covariates |
| AIC       | 655.729              | 645.890                        |
| SC        | 660.083              | 654.598                        |
| -2 Log L  | <b>653.729</b>       | <b>641.890</b>                 |

<sup>21</sup> SAS, STAT User's Guide, Version 8, Volume 2 (Cary, NC: SAS Publishing, 1999), p. 1948.

| Testing Global Null Hypothesis: BETA=0   |            |          |                |            |            |
|--|------------|----------|----------------|------------|------------|
| Test                                     | Chi-Square | DF       | Pr >           | ChiSq      |            |
| Likelihood Ratio                         | 11.8392    | 1        |                | 0.0006     |            |
| Score                                    | 9.7585     | 1        |                | 0.0018     |            |
| Wald                                     | 9.2203     | 1        |                | 0.0024     |            |
| Analysis of Maximum Likelihood Estimates |            |          |                |            |            |
| Parameter                                | DF         | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
| Intercept                                | 1          | -0.7678  | 0.1303         | 34.7133    | <.0001     |
| priors                                   | 1          | -0.0749  | 0.0247         | 9.2203     | 0.0024     |

Looking at Display 9, it is evident from the model fit statistics that  $-2 \ln L(\beta_0) = 653.729$  and  $-2 \ln L(\beta_0, \beta_5) = 641.890$ . In addition, the results for the likelihood ratio test indicate that the new model is significant with  $G = 11.8392$  and the p-value = 0.0006. The new parameter estimates are also given, with the Wald p-value indicating that the variable *priors* is indeed significant to the model. Because the variable *priors* has been entered, SAS rechecks the significance of all entered parameters, in this case only the variable *priors*, to make sure that all of the included variables are still significant to the new model. Next, the likelihood of the seven models containing the intercept, *priors*, and one of the remaining variables is examined. Another variable is entered based upon the calculation of the new Score chi-square statistics.

**Display 10: Analysis of Priors and the Effects not in the Model**

| Analysis of Effects in Model |    |            |            |
|------------------------------|----|------------|------------|
| Effect                       | DF | Chi-Square | Pr > ChiSq |
| priors                       | 1  | 9.2203     | 0.0024     |

| Analysis of Effects Not in the Model |          |               |               |
|--------------------------------------|----------|---------------|---------------|
| Effect                               | DF       | Chi-Square    | Pr > ChiSq    |
| age                                  | 1        | 4.0191        | 0.0450        |
| beck                                 | 1        | 0.3415        | 0.5590        |
| <b>IV1</b>                           | <b>1</b> | <b>5.0971</b> | <b>0.0240</b> |
| IV2                                  | 1        | 0.1514        | 0.6972        |
| race                                 | 1        | 3.4626        | 0.0628        |
| treatmnt                             | 1        | 5.0161        | 0.0251        |
| site                                 | 1        | 0.8312        | 0.3619        |

By the information given in Display 10, *IV1* is now entered into the model, having the largest Score chi-square statistic. The resulting information regarding likelihood values, the likelihood ratio test, and the parameter estimates for the new model is shown in Display 11.

**Display 11: Step 2--IV1 is Entered into the Model**

Model Convergence Status  
Convergence criterion (GCONV=1E-8) satisfied.

| Model Fit Statistics |                |                          |
|----------------------|----------------|--------------------------|
| Criterion            | Intercept Only | Intercept and Covariates |
| AIC                  | 655.729        | 642.718                  |
| SC                   | 660.083        | 655.781                  |
| -2 Log L             | <b>653.729</b> | <b>636.718</b>           |

| Testing Global Null Hypothesis: BETA=0 |                |    |               |
|--|----------------|----|---------------|
| Test                                   | Chi-Square     | DF | Pr > ChiSq    |
| Likelihood Ratio                       | <b>17.0105</b> | 2  | <b>0.0002</b> |
| Score                                  | 15.2994        | 2  | 0.0005        |
| Wald                                   | 14.5859        | 2  | 0.0007        |

| Analysis of Maximum Likelihood Estimates |    |          |                |                 |                  |
|--|----|----------|----------------|-----------------|------------------|
| Parameter                                | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq       |
| Intercept                                | 1  | -0.6460  | 0.1387         | <b>21.6763</b>  | <b>&lt;.0001</b> |
| IV1                                      | 1  | -0.4740  | 0.2108         | <b>5.0552</b>   | <b>0.0246</b>    |
| priors                                   | 1  | -0.0597  | 0.0248         | <b>5.8077</b>   | <b>0.0160</b>    |

This process continues, at each step adding another variable and then using the deviance test to make sure that no variables should be removed from the new model. Once SAS has finished the stepwise procedure, meaning that no additional significant variables remain, a final model is outputted. The final model for our data set is shown in Display 12.

**Display 12: The Stepwise Final Model**

| Step | Effect Entered | Effect Removed | DF | Number In | Score Chi-Square | Wald Chi-Square |
|------|----------------|----------------|----|-----------|------------------|-----------------|
| 1    | priors         |                | 1  | 1         | 9.7585           | .               |
| 2    | IV1            |                | 1  | 2         | 5.0971           | .               |
| 3    | age            |                | 1  | 3         | 6.4484           | .               |
| 4    | treatmnt       |                | 1  | 4         | 4.7295           | .               |
| 5    | IV2            |                | 1  | 5         | 4.8584           | .               |

| Summary of Stepwise Selection |            |                                    |
|-------------------------------|------------|------------------------------------|
| Step                          | Pr > ChiSq | Variable Label                     |
| 1                             | 0.0018     | Prior Drug Treatments              |
| 2                             | 0.0240     | Indicator for Recent IV Drug Use   |
| 3                             | 0.0111     | Age at Enrollment                  |
| 4                             | 0.0296     | Treatment: 0=Short, 1=Long         |
| 5                             | 0.0275     | Indicator for Previous IV Drug Use |

In this example, the five variables found to be significant to the model are *priors*, *IV1*, *age*, *treatmnt*, and *IV2*. Looking at the parameter estimates and p-values given for the model containing these five variables in Display 13, an equation can be estimated for the best model as determined by the stepwise procedure.

**Display 13: Stepwise Final Parameter Estimates**  
Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------|------------|
| Intercept | 1  | -2.3327  | 0.5484         | 18.0956         | <.0001     |
| age       | 1  | 0.0526   | 0.0172         | 9.3378          | 0.0022     |
| IV1       | 1  | -0.8056  | 0.2445         | 10.8542         | 0.0010     |
| IV2       | 1  | -0.6237  | 0.2847         | 4.7989          | 0.0285     |
| priors    | 1  | -0.0637  | 0.0256         | 6.1858          | 0.0129     |
| treatmnt  | 1  | 0.4513   | 0.1986         | 5.1649          | 0.0230     |

Thus,

$$h(\vec{x}_i) = -2.3327 + 0.0526x_{1i} - 0.8056x_{3i} - 0.6237x_{4i} - 0.0637x_{5i} + 0.4513x_{7i}.$$

Again, if other variables are deemed practically important to the model, these could be included as well.

Another popular procedure for fitting these logistic models is the best subset procedure. There are several different criterion used to determine “best” for these computations. SAS uses the Score chi-square statistic, which was discussed in a previous section. In general, whichever model has the highest Score statistic is considered the best model. However, because a degree of freedom is lost for every additional variable added to a model, the best model would not only have a relatively high Score chi-square statistic, but also a small number of predictor variables. Calling up the best subsets procedure in SAS, the two best models for every level of q is shown in the following output, with q representing the number of predictor variables:

**Display 14: Best Subset by the Score Criterion**

| Number of Variables | Score Chi-Square | Variables Included in Model                |
|---------------------|------------------|--|
| 1                   | 9.7585           | priors                                     |
| 1                   | 9.7368           | IV1  |
| 2                   | 15.2994          | IV1 priors                                 |
| 2                   | 14.8924          | priors treatmnt                            |
| 3                   | 21.2275          | age IV1 priors                             |
| 3                   | 20.4581          | age IV1 IV2                                |
| 4                   | 26.1214          | age IV1 IV2 priors                         |
| 4                   | 25.9623          | age IV1 priors treatmnt                    |
| 5                   | 31.1565          | age IV1 IV2 priors treatmnt                |
| 5                   | 27.4079          | age IV1 priors race treatmnt               |
| 6                   | 32.0446          | age IV1 IV2 priors race treatmnt           |
| 6                   | 31.6135          | age IV1 IV2 priors treatmnt site           |
| 7                   | 32.6795          | age IV1 IV2 priors race treatmnt site      |
| 7                   | 32.0481          | age beck IV1 IV2 priors race treatmnt      |
| 8                   | 32.6798          | age beck IV1 IV2 priors race treatmnt site |

Looking at the above output, we see that the model with all eight variables provides the highest Score statistic, as would be expected. However, there are several other models with fewer predictor variables that have Score statistics almost as high as this model. Looking at the subsets for the models with five predictor variables, we see that the first model has a Score statistic of 31.1565, only slightly smaller than the maximum value. In addition, it only has five variables. Therefore this model, which is the same model selected by our stepwise method, would be a good choice for the best subset of predictor variables. In this case, the model decision based upon the best subset method is equivalent to the model decision in the stepwise procedure, but it should be noted that these procedures will not always produce the same results.

### Adequacy of the Model:

By means of both methods, the strongest model with which to fit the data is the following logit model:

$$h(\vec{x}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{3i} + \beta_3 x_{4i} + \beta_4 x_{5i} + \beta_5 x_{7i}.$$

After fitting this model, a few diagnostic measures can be used to look at the fit of the model.

The estimated model is as follows:

**Display 15: Final Model**

| Parameter | DF | Estimate | Standard Error | Chi-Square | Wald    | Pr > ChiSq |
|-----------|----|----------|----------------|------------|---------|------------|
| Intercept | 1  | -2.3327  | 0.5484         | 18.0956    | 18.0956 | <.0001     |
| age       | 1  | 0.0526   | 0.0172         | 9.3378     | 9.3378  | 0.0022     |
| IV1       | 1  | -0.8056  | 0.2445         | 10.8542    | 10.8542 | 0.0010     |
| IV2       | 1  | -0.6237  | 0.2847         | 4.7989     | 4.7989  | 0.0285     |
| priors    | 1  | -0.0637  | 0.0256         | 6.1858     | 6.1858  | 0.0129     |
| treatmnt  | 1  | 0.4513   | 0.1986         | 5.1649     | 5.1649  | 0.0230     |

Odds Ratio Estimates

| Effect   | Point Estimate | 95% Wald Confidence Limits |       |
|----------|----------------|----------------------------|-------|
| age      | 1.054          | 1.019                      | 1.090 |
| IV1      | 0.447          | 0.277                      | 0.722 |
| IV2      | 0.536          | 0.307                      | 0.936 |
| priors   | 0.938          | 0.892                      | 0.987 |
| treatmnt | 1.570          | 1.064                      | 2.318 |

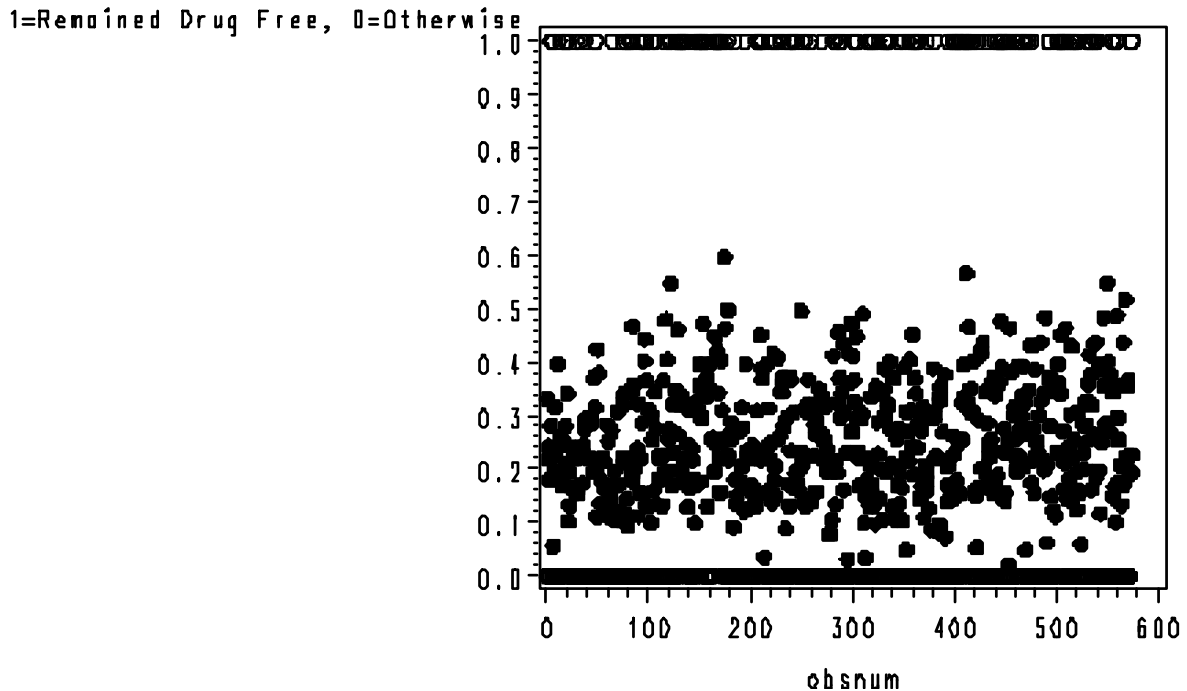
Therefore, we again find that

$$h(\vec{x}_i) = -2.3327 + 0.0526x_{1i} - 0.8056x_{3i} - 0.6237x_{4i} - 0.0637x_{5i} + 0.4513x_{7i}.$$

By looking at the odds ratio estimates in Display 15, we see that an additional year in age increases the probability of remaining drug free by 5.4%. Recent IV drug use decreases that probability by 55.3% while previous IV drug use decreases that probability by 46.4%. We also can note that for every additional prior drug treatment to which the subject was exposed, the probability of remaining drug free falls by 6.2%. Finally, undergoing the long treatment process as opposed to the short treatment process increases the chances of remaining drug free by 57%.

A graphic image of this fit is depicted below:

**Figure 2: Final Logistic Fit**



One way of assessing the fit of this model is by means of the Hosmer-Lemeshow Goodness of Fit test. In order to perform this test, the data must be divided into  $g$  groups, where it is normally recommended that  $g$  is equal to ten. For all of the goodness-of-fit tests, the hypotheses are:

$$H_0 : E\{Y\} = \left(1 + e^{-\beta_0 - \beta_1 x_{i1} - \dots - \beta_5 x_{i5}}\right)^{-1}, \text{ or the model fit is appropriate}$$

$$H_A : E\{Y\} \neq \left(1 + e^{-\beta_0 - \beta_1 x_{i1} - \dots - \beta_5 x_{i5}}\right)^{-1}, \text{ or the model fit is inappropriate.}$$

The associated test statistic is:

$$\hat{C} = \sum_{k=1}^g \left\{ \frac{\left[ \sum y_{ik} - n_k \bar{\pi}_k(x_i) \right]^2}{n_k \bar{\pi}_k(x_i) [1 - \bar{\pi}_k(x_i)]} \right\}$$

where  $\hat{C}$  is a chi-square random variable with  $(g - 2)$  degrees of freedom. Therefore, the null hypothesis will be rejected if  $\hat{C} > \chi^2(1 - \alpha; g - 2)$  while the null hypothesis will not be rejected if  $\hat{C} \leq \chi^2(1 - \alpha; g - 2)$ .<sup>22</sup> SAS performs this goodness of fit test using  $g = 10$ . According to SAS,

**Display 16: Hosmer and Lemeshow Goodness of Fit Test**

| Group | Total | drugFree = 1 |          | drugFree = 0 |          |
|-------|-------|--------------|----------|--------------|----------|
|       |       | Observed     | Expected | Observed     | Expected |
| 1     | 59    | 6            | 5.87     | 53           | 53.13    |
| 2     | 59    | 8            | 8.80     | 51           | 50.20    |
| 3     | 58    | 9            | 10.19    | 49           | 47.81    |
| 4     | 59    | 12           | 12.25    | 47           | 46.75    |
| 5     | 58    | 13           | 13.40    | 45           | 44.60    |
| 6     | 58    | 17           | 15.25    | 41           | 42.75    |
| 7     | 58    | 21           | 17.30    | 37           | 40.70    |
| 8     | 58    | 20           | 19.34    | 38           | 38.66    |
| 9     | 59    | 18           | 22.22    | 41           | 36.78    |
| 10    | 49    | 23           | 22.39    | 26           | 26.61    |

| Hosmer and Lemeshow Goodness-of-Fit Test |               |
|--|---------------|
| Chi-Square                               | Pr > ChiSq    |
| <b>3.0300</b>                            | <b>0.9325</b> |

<sup>22</sup> Ryan, pp. 278-279.

With  $\hat{C} = 3.03$  and a p-value of 0.9325, the null hypothesis will not be rejected, resulting in the conclusion that, according to the Hosmer and Lemeshow Goodness-of-Fit test, the model does in fact provide a good fit to the data.

A similar goodness-of-fit test can be performed on the same set of hypotheses as above, where in this case the test would be based upon the model deviances. The test statistic here would be the deviance of the model that has been fit, as defined above, or

$$D = -2 \sum_{i=1}^n \left[ y_i \ln \left( \frac{\hat{\pi}(\bar{x}_i)}{y_i} \right) + (1 - y_i) \ln \left( \frac{1 - \hat{\pi}(\bar{x}_i)}{1 - y_i} \right) \right]$$

where

$$\hat{\pi}(\bar{x}_i) = \frac{e^{b_0 + b_1 x_{i1} + \dots + b_p x_{ip}}}{1 + e^{b_0 + b_1 x_{i1} + \dots + b_p x_{ip}}}$$

The null hypothesis should be rejected if  $DEV(\bar{x}_i) > \chi^2(1 - \alpha; n - p + 1)$  and the null hypothesis should not be rejected if  $DEV(\bar{x}_i) \leq \chi^2(1 - \alpha; n - p + 1)$ .<sup>23</sup> Given the drug data, the deviance is listed as 620.587 while  $\chi^2(1 - 0.05; 575 - 6) = 625.6015$ .

| Model Fit Statistics |                |                          |
|----------------------|----------------|--------------------------|
| Criterion            | Intercept Only | Intercept and Covariates |
| AIC                  | 655.729        | 632.587                  |
| SC                   | 660.083        | 658.713                  |
| -2 Log L             | 653.729        | <b>620.587</b>           |

Because the calculated deviance is less than the chi-squared critical value, once again the null hypothesis will not be rejected, further supporting the adequacy of the fit of this model to the data.

## Conclusion

---

<sup>23</sup> Neter, p. 595.

In the modeling of a binary response variable, the logistic regression technique serves as a key element to any analysis. By the very nature of the response variable, the results of the fitted model can easily be interpreted in terms of the probability that the event of interest occurs. The model fitting procedure of maximum likelihood estimation is fairly computationally intensive, but with so many statistical programs available, the results can be easily computed. This technique is very important for statistical analyses, particularly in areas such as health science where until the development of logistic regression, there was no way of accurately fitting a model to the data. With the current focus in medicine lying toward finding treatments and even cures for many of the health issues plaguing our society, such as cancer, AIDS, and even the common cold, logistic regression will serve as a good tool for determining whether or not these proposed treatments are effective. The growing desire of the general human population to gain more knowledge in medicinal and psychological areas, as well as many other areas, will cause the continued use of this regression technique.

## SAS Code Used for Analysis

```
OPTIONS LS=75;
FILENAME indata 'H:\Comps\drugData.dat';
DATA drugUse;
  INFILE indata;
  INPUT ID age beck IV priors race treatmnt site drugFree;
  obsnum=_N_;
LABEL
  ID=Identification Code
  age=Age at Enrollment
  beck=Beck Depression Score
  IV=IV Drug Use History
  priors=Prior Drug Treatments
  race='Race: 0=White, 1=Other'
  treatmnt='Treatment: 0=Short, 1=Long'
  site='Site: 0=A, 1=B'
  drugFree='1=Remained Drug Free, 0=Otherwise'
;
DATA drugUse2;
  SET drugUse;
  IF IV=3 THEN IV1=1;
  ELSE IV1=0;
  IF IV=2 THEN IV2=1;
  ELSE IV2=0;
LABEL
  IV1=Indicator for Recent IV Drug Use
  IV2=Indicator for Previous IV Drug Use
;
/*Simple Logistic Regression Model*/
PROC LOGISTIC DATA=drugUse2 descending;
  MODEL drugFree=treatmnt/link=logit;
  OUTPUT OUT=simple P=yhat RESDEV=devResid;
RUN;
PROC GPLOT DATA=simple;
  PLOT drugFree*treatmnt yhat*treatmnt/overlay;
  symbol1 v=circle line=none c=black;
  symbol2 v=dot line=spline c=black;
/*Multiple Logistic Regression Model*/
PROC LOGISTIC DATA=drugUse2 descending;
  MODEL drugFree=age race treatmnt/
  link=logit Rsquare;
  OUTPUT OUT=multiple P=yhat RESDEV=devResid;
RUN;
/*Model Building*/
/*Stepwise*/
PROC LOGISTIC DATA=drugUse2 descending;
  MODEL drugFree=age beck IV1 IV2 priors race treatmnt site/
  link=logit selection=stepwise details slentry=0.20 slstay=0.20;
  OUTPUT OUT=stepwise P=yhat;
RUN;
/*Model Building*/
/*Best Subset*/
PROC LOGISTIC DATA=drugUse2 descending;
  MODEL drugFree=age beck IV1 IV2 priors race treatmnt site/
  selection=score details;
  OUTPUT OUT=best P=yhat;
RUN;
/*Model Building*/
/*Limited Best Subset*/
PROC LOGISTIC DATA=drugUse2 descending;
  MODEL drugFree=age beck IV1 IV2 priors race treatmnt site/
  selection=score best=2 details;
  OUTPUT OUT=topbest P=yhat;
RUN;
/*Temporary Model*/
PROC LOGISTIC DATA=drugUse2 descending;
  MODEL drugFree=age IV1 IV2 priors treatmnt/Rsquare Influence Lackfit
  Iplots;
  OUTPUT OUT=temp P=yhat RESDEV=devResid RESCHI=chiResid
  H=hat DIFDEV=devChng DIFCHISQ=chiChng DFBETAS=_ALL_;
RUN;
PROC GPLOT DATA=temp;
  PLOT drugFree*obsnum yhat*obsnum/overlay;
  symbol1 v=circle i=none c=black;
  symbol2 v=dot i=none c=black;
RUN;
```

## Sources

- Chatterjee, Samprit, Ali S. Hadi, and Bertram Price, *Regression Analysis by Example* (New York: John Wiley & Sons, 2000)
- Hogg, Robert V., and Allen T. Craig, *Introduction to Mathematical Statistics* (London: The MacMillan Company, 1970)
- Hosmer, David W., and Stanley Lemeshow, *Applied Logistic Regression* (New York: John Wiley & Sons, 1989)
- Hosmer, David W., and Stanley Lemeshow, *Applied Logistic Regression* (New York: John Wiley & Sons, 1989)
- Menard, Scott, *Applied Logistic Regression Analysis* (Thousand Oaks, California: Sage Publications, 1995)
- Neter, John, William Wasserman, Christopher J. Nachtsheim, and Michael H. Kutner, *Applied Linear Regression Models*, 3<sup>rd</sup> ed. (Chicago: McGraw-Hill, 1996)
- Ryan, Thomas P., *Modern Regression Methods* (New York: John Wiley & Sons, Inc., 1997)
- SAS, STAT User's Guide, Version 8, Volume 2 (Cary, NC: SAS Publishing, 1999).
- SAS version 8.2