Some Issues in Using PROC LOGISTIC for Binary Logistic Regression

by David C. Schlotzhauer

Contents

Abstract

For regression situations in which the response is binary (i.e. it has only two levels), logistic regression is often used. Proper interpretation of the estimated parameters of the logistic model depends on the ordering of the two response levels. Similarly, the construction and interpretation of odds ratio estimates (available in Release 6.07 TS301 and later) are affected by response level ordering. In addition to these issues, this article discusses using the fitted model to compute predicted probabilities. By selecting a cutoff value, these probabilities can be used to classify observations into one of the response levels. This leads to construction of a classification table summarizing the predicted and observed responses. The fitted model and cutoff value can also be used to classify future observations.

1. The Effect of Response Level Ordering on Parameter Estimate Interpretation

For binary response data, it is important to remember that the default response function modeled by the LOGISTIC procedure is

$$\text{logit}(p) \equiv \log\left(\frac{p}{1-p}\right)$$

where p is the probability of the response level identified in the Response Profiles section as having Ordered Value 1. In the binary response case, the response levels can be ordered in two ways: the event of interest as Ordered Value 1 and the nonevent as Ordered Value 2, or the event of interest as Ordered Value 2 and the nonevent as Ordered Value 1. Either ordering can be used, but correct interpretation of the results must take this into account as we'll see below.

Consider a model with a single explanatory variable, X,

$$\text{logit}(p) = \alpha + \beta x \qquad (1)$$

For positive β, as X increases, logit(p) increases. But it is also true that logit(p) monotonically increases with p, meaning that logit(p) never decreases as p increases as shown in Figure 1 below.

Since p and logit(p) increase and decrease together, if β is positive, increasing X increases p. Similarly, a negative β indicates that increasing X decreases p. If the event of interest is Ordered Value 1, then p is the probability of an event. Consequently, by simply examining the sign of an explanatory variable's parameter estimate, you can easily determine the effect of the variable on the

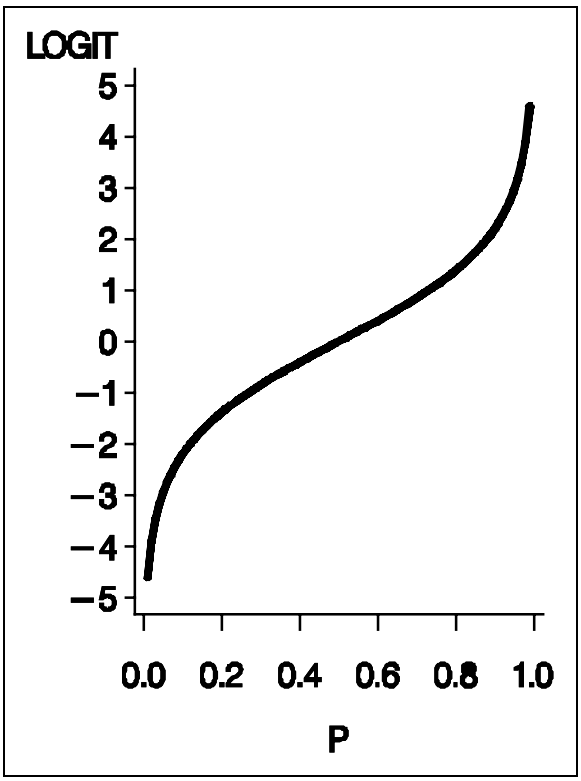

Figure 1. Logit(p) increases with p

probability of an event. If the event of interest is not Ordered Value 1, and you do not take this into account, then the parameter signs may seem backwards to you. It is very important, therefore, to know which of your response levels is Ordered Value 1 before looking at the remainder of the analysis output.

For instance, if the response levels are ordered such that the event of interest is Ordered Value 2, then p becomes the probability of the nonevent (since it is Ordered Value 1), and, when β is positive, the probability of a nonevent increases as X increases. In this case, if you interpret the positive β as meaning that the probability of the event increases as X increases, you would be assuming, incorrectly, that the event of interest is Ordered Value 1.

The importance of knowing the response level ordering also applies to binary probit analysis done either through the LOGISTIC procedure with the LINK=NORMIT option or through the PROBIT procedure. Just as with logistic regression, the parameter estimates in a probit model may be interpreted more intuitively if the event of interest is the first ordered level.

The following example uses the common convention of numerically coding the response levels as 0 and 1 with the value 1 corresponding to the event of interest. If the possible responses are DISEASE and NO DISEASE, with DISEASE being the event of interest, then a response variable is often defined as follows:

$$Y = \begin{cases} 1 \textit{ if } \text{DISEASE} \\ 0 \textit{ if } \text{NO DISEASE} \end{cases}$$

Similarly, an explanatory variable, EXPOSURE, might be defined like this:

$$\text{EXPOSURE} = \begin{cases} 1 \textit{ if } \text{EXPOSED} \\ 0 \textit{ if } \text{NOT EXPOSED} \end{cases}$$

Suppose the data to be modeled are summarized in the following table:

EXPOSURE	Y		
Frequency| Row Pct	0	1	Total
0	45 81.82	10 18.18	55
1	5 11.11	40 88.89	45
Total	50	50	100

Figure 2. Association of Exposure and Disease

Notice in Figure 2 that the EXPOSED group (the last row) is much more likely to be diseased than the NOT EXPOSED group (the first row). The selected sections of LOGISTIC output shown in Output 1 are generated by

submitting the following statements[1]:

```
data disease;
   input y exposure freq;
   cards;
     0 0 45
     0 1 5
     1 0 10
     1 1 40
   ;
proc logistic data=disease;
   model y = exposure;
   freq freq;
   run;
```

order. The ordering scheme may be altered by the use of the ORDER= and DESCENDING options on the PROC LOGISTIC statement as needed. Because of the way that we've coded our response, Y, the default ordering results in Y=0 as Ordered Value 1 and Y=1 as Ordered Value 2.

Notice how this affects parameter estimate interpretation. The negative parameter estimate for EXPOSURE (-3.5835) does *not* mean that as EXPOSURE increases from 0 to 1, Y tends to decrease from 1 (DISEASE) to 0 (NO DISEASE). Since Ordered Value 1

Response Profile

Ordered Value	Y	Count
1	0	50
2	1	50

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	1.5041	0.3496	18.5093	0.0001	.	4.500
EXPOSURE	1	-3.5835	0.5893	36.9839	0.0001	-0.987849	0.028

Output 1. Model with p = Pr(No Disease)

The first thing to notice is the level-ordering displayed in the Response Profile. Note that Ordered Value 1 corresponds to NO DISEASE (Y=0) rather than the event of interest, DISEASE. By default, LOGISTIC orders the response values (or the associated formatted values if defined) in increasing alphanumeric

corresponds to Y=0 (NO DISEASE), p is defined to be the probability of a nonevent. The negative parameter estimate for EXPOSURE actually indicates that as EXPOSURE increases from 0 to 1, logit(p) and p decrease, meaning the probability of NO DISEASE decreases. Since the probabilities of DISEASE and NO DISEASE must sum to one, the probability of DISEASE increases as the probability of NO DISEASE decreases. So, the negative EXPOSURE parameter means that as EXPOSURE increases, the probability of DISEASE increases. This is consistent with the data as seen in the Figure 2 above – the probability of DISEASE increases from 18% in the EXPOSURE=0 group to 89% in the EXPOSURE=1 group.

[1] The FREQ statement, DESCENDING option, odds ratios with optional confidence limits and multiple-cutoff classification table illustrated in this article are features available in Release 6.07 TS301 and later. These features are described in **Technical Report P-229 SAS/STAT Software: Changes and Enhancements, Release 6.07**. All output in this article was generated using Release 6.07 TS301.

Though we finally arrived at the correct interpretation by taking the level ordering into account, it was difficult simply because all information printed by LOGISTIC refers to Ordered Value 1 (NO DISEASE) but we wanted to make statements about Ordered Value 2 (DISEASE). If you ensure that the event of interest is Ordered Value 1, then interpretation is much easier. For instance, if you reverse the response level ordering so that p becomes the probability of DISEASE, the parameter estimates change sign, as seen below in Output 2.

The positive EXPOSURE parameter now has the intuitive interpretation of an increase in EXPOSURE yielding an increase in the probability of DISEASE[2].

There are several ways that you can reverse the response level ordering to yield Output 2 rather than Output 1:

- The simplest way, available in Release 6.07 TS301 and later, uses the option, DESCENDING. Specify the DESCENDING option on the PROC LOGISTIC statement to reverse the default ordering of Y from 0,1 to 1,0 making 1 (DISEASE) the level with Ordered Value 1:

```
proc logistic data=disease
    descending;
    model y = exposure;
    freq freq;
    run;
```

- Assign a format to Y such that the first formatted value (when the formatted values are put in sorted order) corresponds to DISEASE. For this example, Y = 0 could be assigned formatted value 'no disease' and Y = 1 assigned formatted value 'disease'.

```
proc format;
    value disfmt 1='disease'
    0='no disease';
    run;
proc logistic data=disease;
    model y = exposure;
    freq freq;
    format y disfmt.;
    run;
```

- Sort the input data so that observations with Y = 1 occur before observations with Y = 0, then use the ORDER=DATA option on the PROC LOGISTIC statement. For this example, sort the data before running LOGISTIC with the following statements:

```
proc sort data=disease;
    by descending y;
    run;
proc logistic data=disease
    order=data;
    model y = exposure;
    freq freq;
    run;
```

- Create a new variable to replace Y as the response variable in the MODEL statement of LOGISTIC such that observations with Y = 1 take on the first value of the new variable (when put in sorted order).

```
data disease2;
    set disease;
    if y = 0 then
       y1 = 'no disease';
    else y1 = 'disease';
    run;
proc logistic data=disease2;
    model y1 = exposure;
    freq freq;
    run;
```

- Create a new variable (N, say) with constant value 1 for each observation. Use the events/trials MODEL statement syntax with Y as the event variable and N as the trial variable. (Note that this method depends on Y having values 0 and 1, where 1 represents the event of interest). When this is done, the ratio y/n for each observation is either 0/1, indicating no event occurred in this single trial, or 1/1 indicating an event did occur in this trial.

---

[2] If you have used the Version 5 contributed procedure, LOGIST, for binary logistic regression, you noticed that it requires a 0,1 coding of the response variable (the response variable in LOGISTIC can have any numeric or character values) and it always defines p as the probability of level 1. For this example, the parameter estimates obtained by LOGIST would match those in Output 2.

Response Profile

Ordered Value	Y	Count
1	1	50
2	0	50

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.5041	0.3496	18.5093	0.0001	.	0.222
EXPOSURE	1	3.5835	0.5893	36.9839	0.0001	0.987849	36.000

Output 2. Model with p = Pr(Disease)

```
data disease3;
   set disease;
   n=1;
   run;
proc logistic data=disease3;
   model y/n = exposure;
   freq freq;
   run;
```

2. Odds Ratios

The odds of an event is the ratio of the probability of an event (p) to the probability of a nonevent (1-p):

$$\text{odds} = \frac{p}{1-p}$$

For instance, in Figure 2, the odds of disease in the exposed group are 88.89/11.11 = 8 meaning that disease is 8 times more likely than no disease in the exposed group. For the unexposed group, the odds are 0.222.

An explanatory variable's odds ratio, given in the last column of the LOGISTIC procedure output, tells you how the odds of an event change as you increase the variable by one[3]. If you compute the odds at two different settings of the explanatory variable, the odds ratio is just the ratio of these two odds. In Figure 2, the odds ratio is 8/0.222 = 36 meaning that the odds of disease in the exposed group are 36 times the odds of disease in the unexposed group.

Notice that logit(p) is just the log odds:

$$\text{logit}(p) \equiv \log(p/(1-p)) \equiv \text{log odds}$$

Then, using Model 1 with estimated parameters a and b,

$$\text{odds} = \exp(a+bx) \quad .$$

The odds ratio comparing the odds for two values of X, x_1 and x_2 (where $x_1 < x_2$), is

$$\begin{aligned} \text{odds ratio} &= \frac{\exp(a+bx_2)}{\exp(a+bx_1)} \\ &= \exp(b(x_2-x_1)) \\ &= (\exp(b))^u \quad , \end{aligned}$$

where $u = x_2 - x_1$. When X is a binary variable

[3] You can also request a 100(1-α)% confidence interval for the odds ratio by specifying the RISKLIMITS and ALPHA=α options on the MODEL statement of the LOGISTIC procedure.

such that $x_2 - x_1 = 1$, then the odds ratio estimate is simply exp(b). In our example, the explanatory variable, EXPOSURE, meets this condition.

The effect of response level ordering on the interpretation of odds ratios for binary explanatory variables is discussed in sections 2.1 and 2.2. Odds ratios for continuous explanatory variables are covered in section 2.3.

2.1 Binary Explanatory Variable – Modeling the Event

Just as in parameter estimate interpretation, the interpretation of odds ratios requires that you be aware of the response level ordering. In the disease-exposure example, when DISEASE is Ordered Value 1, so that p is the probability of DISEASE as in Output 2 above, the fitted Model 1 is

$$\begin{aligned} \text{logit}(p_x) &= \log\left(\frac{p_x}{1-p_x}\right) \\ &= -1.5041 + 3.5835 \cdot 1 \\ &= 2.0794 \end{aligned} \qquad (2)$$

when EXPOSURE=1 and

$$\begin{aligned} \log\left(\frac{p_{nx}}{1-p_{nx}}\right) &= -1.5041 + 3.5835 \cdot 0 \\ &= -1.5041 \end{aligned} \qquad (3)$$

when EXPOSURE=0, where p_x represents the probability of disease in the exposed group and p_{nx} is the probability of disease in the unexposed group. Subtracting Equation 3 from Equation 2 gives

$$\log\left(\frac{p_x}{1-p_x}\right) - \log\left(\frac{p_{nx}}{1-p_{nx}}\right) = 3.5835 \quad . \qquad (4)$$

Equation 4 shows that the parameter estimate for EXPOSURE is the difference in log odds for the two groups. Writing the difference in logs as the log of the ratio and exponentiating both sides of Equation 4 gives

$$\frac{p_x/(1-p_x)}{p_{nx}/(1-p_{nx})} = e^{3.5835} = 36 \quad . \qquad (5)$$

Notice that the left-hand side is the odds ratio and 36 is the value shown in Output 2 above. Notice that an algebraically-equivalent statement is that the odds of disease in the unexposed group are 1/36 = 0.028 times the odds of disease in the exposed group.

2.2 Binary Explanatory Variable – Modeling the Nonevent

Let's see what happens if we model the nonevent, NO DISEASE. When NO DISEASE is Ordered Value 1 as in Output 1 above, let p' represent the probability of the nonevent, NO DISEASE. The fitted Model 1 becomes

$$\text{logit}(p_x') = \log\left(\frac{p_x'}{1-p_x'}\right) = 1.5041 - 3.5835$$

when EXPOSURE=1 and

$$\log\left(\frac{p'_{nx}}{1-p'_{nx}}\right) = 1.5041$$

when EXPOSURE=0. The odds ratio (as shown in Output 1) is

$$\frac{p_x'/(1-p_x')}{p'_{nx}/(1-p'_{nx})} = e^{-3.5835} = 0.028 \quad .$$

The odds of being undiseased in the exposed group are 0.028 times those in the unexposed group. So, the exposed group has a much lower odds of being undiseased. Notice that this odds ratio is just the reciprocal of the previous odds ratio, 1/36, and it could have been found from Equation 5 by noting that p'=1-p and using a little algebra. Generally, changing the order of the columns of Figure 2 (i.e., changing the response level ordering) changes the sign of the parameter estimate and inverts the odds ratio as we've just shown. Inversion of the odds ratio also occurs if you

switch the rows of Figure 2, as shown at the end of Section 2.1.

2.3 Continuous Explanatory Variable

For a continuous variable, X, exp(b) represents the change in odds for a unit increase in X. LOGISTIC always estimates the odds ratio with exp(b) since it treats all explanatory variables as continuous. To obtain an estimate of the odds ratio for a change of *u* units, simply raise the odds ratio reported by LOGISTIC to the power *u*. This will also allow you to obtain an estimate of the odds ratio if you have a binary variable whose values differ by *u* units.

To demonstrate the continuous explanatory variable case, suppose that, in a separate group of individuals, age is used to predict disease. In this group, 30 of 100 individuals have the disease. A subset of the data is shown below in Figure 3.

AGE	Y
32	0
79	1
44	0
36	0
76	1
79	1
53	0
52	0
23	0
24	0

Figure 3. Subset of Age and Disease Data

The following statements:

```
proc logistic data=age
   descending;
   model y = age;
   run;
```

fit a logistic model with Y=1 (DISEASE) as Ordered Value 1 and produce Output 3 below.

The odds ratio of 1.264 indicates that the odds of disease increase by a factor of 1.264 for each increase in age of one year. For an increase of 10 years the odds of disease increase by a factor of 1.264^{10} =10.4 .

3. Predicted Probabilities

As in simple regression, predicted values can be obtained from a logistic regression model. However, the predicted value from a binary logistic model is the estimated probability of an observation being an event (which is Ordered Value 1 when you use LOGISTIC). In our disease-exposure example, Model 1 becomes Equation 2 when EXPOSURE=1, yielding a predicted logit of 2.0794. To obtain the estimate of p, the probability of an event, solve the estimated Model 1 for p:

$$\text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = a + bx$$

$$\frac{\hat{p}}{1-\hat{p}} = \exp(a + bx)$$

Multiplying both sides by 1-$\hat{p}$ and isolating $\hat{p}$ gives

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-13.8651	3.0808	20.2538	0.0001	.	0.000
AGE	1	0.2344	0.0532	19.3880	0.0001	2.167008	1.264

Output 3. Model for Disease and Age

$$\hat{p} = \frac{\exp(a+bx)}{1+\exp(a+bx)} \quad .$$

Multiplying top and bottom by 1/exp(a + bx) gives

$$\hat{p} = \frac{1}{1+\exp(-a-bx)} \quad .$$

When EXPOSURE=1, the predicted probability of disease is 1/(1+exp(-2.0794)) = 0.8889. The predicted probability of no disease is $\hat{p}' = 1-\hat{p} =$ 0.1111. You can use LOGISTIC to compute and output predicted probabilities for each observation in your data set by using the PREDICTED= option on the OUTPUT statement.

These statements

```
proc logistic data=disease
   descending;
   model y = exposure;
   freq freq;
   output out=probs
          predicted=phat;
   run;
```

create an output data set called PROBS that is a copy of the input data set, but with two additional variables: _LEVEL_ and PHAT. For each observation, PHAT contains the predicted probability of that observation being in the response level shown by _LEVEL_. For binary logistic regression, this level will always be the level with Ordered Value 1. By using the DESCENDING option, we selected Y=1 (DISEASE) to be Ordered Value 1. Output 4 shows the PROBS data set.

Y	EXPOSURE	FREQ	_LEVEL_	PHAT
0	0	45	1	0.18182
0	1	5	1	0.88889
1	0	10	1	0.18182
1	1	40	1	0.88889

Output 4. Predicted Probabilities

4. Predicted by Observed Classification Tables

Using the predicted probabilities, you can use the fitted model to predict whether an observation is an event or a nonevent. It is then possible to construct a table summarizing the predicted and observed responses. The resulting 2×2 table is called a Classification Table, and is one tool that can be used to assess the predictive accuracy of the model.

4.1 Classification Using Predicted Probabilities

To obtain a predicted response for each observation, you apply a decision rule to the predicted probabilities. For instance, you might use a rule that classifies an observation as diseased if $\hat{p}>\hat{p}'$ and as undiseased otherwise. Notice that this rule can be written as $\hat{p}>0.5$. Using the disease-age example, we can add a variable giving the predicted classification for each observation with a simple DATA step as follows:

```
proc logistic data=age
   descending;
   model y = age;
   output out=probs
          predicted=phat;
   run;
data probs;
   set probs;
   pred_dis=0;
   if phat>0.5 then pred_dis=1;
   run;
```

Using this predicted classification variable, PRED_DIS, and the actual response variable, Y, we can create a table showing how many observations were correctly and incorrectly classified by the model:

```
proc freq data=probs;
   tables y*pred_dis /
          norow nocol nopercent;
   run;
```

Output 5 shows the table.

Y	PRED_DIS		
Frequency	0	1	Total
0	64	6	70
1	7	23	30
Total	71	29	100

Output 5. Classification Summary Table

Correctly-classified observations appear on the main diagonal of the table. Using our decision rule's cutoff value of 0.5, the model correctly classified (64 + 23)/100 = 0.87 or 87% of the observations.

4.2 Classification Using Bias-adjusted Predicted Probabilities

LOGISTIC can create a classification table similar to that in Output 5 with the CTABLE and PPROB= options. CTABLE requests that a classification table be created and PPROB= allows you to specify a cutoff value (or more than one cutoff value in release 6.07 TS301 and later).

However, there is an important difference in the two tables. Notice that the table in Output 5 resulted from using all observations to fit the model. Consequently, each observation influenced the model used to classify itself. This can bias the results. Ideally, for a given observation, you would fit a model excluding that observation from the data and then classify the observation using the resulting model. This method could obviously be very expensive if the data set were large. The method used by the CTABLE option is one that approximates this unbiased method and is less expensive. The parameters of the classifying model for each observation are adjusted using the method described in the "Calculation Method" Details section of the LOGISTIC documentation. The observations are then classified according to the cutoff(s) specified in the PPROB= option and the classification table is produced.

The following code produces a classification table for ten different cutoffs: $\hat{p} \geq 0.05$, $\hat{p} \geq 0.10$, ... , $\hat{p} \geq 0.50$:

```
proc logistic data=age
  descending;
  model y = age / ctable
  pprob=(0.05 to 0.5 by 0.05);
  run;
```

The table is displayed in Output 6 below.

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.050	30	47	23	0	77.0	100.0	67.1	43.4	0.0
0.100	30	53	17	0	83.0	100.0	75.7	36.2	0.0
0.150	30	55	15	0	85.0	100.0	78.6	33.3	0.0
0.200	30	60	10	0	90.0	100.0	85.7	25.0	0.0
0.250	29	61	9	1	90.0	96.7	87.1	23.7	1.6
0.300	25	62	8	5	87.0	83.3	88.6	24.2	7.5
0.350	23	62	8	7	85.0	76.7	88.6	25.8	10.1
0.400	23	63	7	7	86.0	76.7	90.0	23.3	10.0
0.450	23	63	7	7	86.0	76.7	90.0	23.3	10.0
0.500	23	63	7	7	86.0	76.7	90.0	23.3	10.0

Output 6. Bias-adjusted Classification Table

Event in this table always refers to Ordered Value 1 – DISEASE in this example. The two columns labeled **Correct** give the number of correctly classified events and nonevents. The **Incorrect Event** column gives the number of nonevents incorrectly classified as events and the **Incorrect Nonevent** column gives the number of events incorrectly classified as nonevents.

The **Percentage Correct** column shows that a $\hat{\rho}$ cutoff in the 0.20 to 0.25 range results in the maximum correct classification rate of 90%. For the 0.20 cutoff, the **Correct Event** column shows that all 30 diseased individuals were correctly classified. However, the 0.25 cutoff incorrectly classifies one diseased individual as undiseased for a false negative rate of 1/62 = 1.6%. When 0.20 is used as the cutoff, the **Incorrect Event** column shows that 10 of the 70 undiseased individuals are incorrectly classified as diseased for a false positive rate of 10/40 = 25%. When 0.25 is used as the cutoff, 9 undiseased individuals are incorrectly classified as diseased for a false positive rate of 9/38 = 23.7%. Your selection of a cutoff may not be merely a choice of the maximum correct classification rate. Depending on the relative costs of false positives and false negatives, you may opt to use a cutoff that gives a slightly lower correct classification rate in order to minimize cost.

Notice when using the 0.5 cutoff, the percentage correct is 86% compared to the 87% found with the previous method that does not adjust for bias.

5. Classifying New Observations

Bias is not a problem if the observations that you wish to classify are not those used to fit the model. You can use LOGISTIC to classify a separate set of observations using a previously constructed model. For instance, suppose you wish to know how well the disease-age model, using the $\hat{\rho}$>=0.25 decision rule, would classify a set of ten new individuals. You can do this by creating a new data set (called AGE2) containing these observations, and, optionally, their actual responses. Then use the following statements to set the response variable to missing. If you have the actual responses for these observations, create a separate variable (called YACTUAL) to hold that information.

```
data unknown;
   set age2;
   yactual=y;
   y=.;
   run;
```

Then, add these observations to the original disease-age data set (AGE) to create a new data set:

```
data both;
   set unknown age;
   run;
```

If you input data set BOTH to LOGISTIC, the procedure ignores the first ten observations, because it always ignores observations which contain missing values. As a result, with the following statements, LOGISTIC uses exactly the same observations as was used originally and produces exactly the same results as in Output 3 above.

```
proc logistic data=both
  descending;
  model y = age;
  output out=probs
         predicted=phat;
  run;
```

However, the output data set (PROBS) created by the OUTPUT statement contains predicted probabilities for *all* observations in the input data set *including* the first ten. Figure 4 shows our ten new observations as they appear in data sets AGE2, UNKNOWN and PROBS.

Now, proceed exactly as in Section 4.1:

```
data probs;
  set probs;
  pred_dis=0;
  if phat>0.25 then pred_dis=1;
  run;
```

If you did not have the actual response values for these individuals, you could still print the first ten observations of data set PROBS to see the predicted classification for each individual

- AGE2 -		------ UNKNOWN -----			--------- PROBS ---------		
AGE	Y	AGE	YACTUAL	Y	AGE	YACTUAL	PHAT
69	1	69	1	.	69	1	0.90963
61	1	61	1	.	61	1	0.60679
77	1	77	1	.	77	1	0.98500
50	0	50	0	.	50	0	0.10483
78	1	78	1	.	78	1	0.98809
45	0	45	0	.	45	0	0.03500
40	0	40	0	.	40	0	0.01111
47	0	47	0	.	47	0	0.05479
31	0	31	0	.	31	0	0.00136
78	1	78	1	.	78	1	0.98809

Figure 4. Data sets leading to predicted probabilities

as shown at the right of Figure 4. Since we do have the actual response values, we can continue with these statements:

```
proc freq data=probs;
  tables yactual*pred_dis /
         norow nocol nopercent;
  run;
```

to produce a classification table for the ten new observations as shown in Output 7 below.

```
YACTUAL     PRED_DIS

Frequency|       0|       1|  Total
---------+--------+--------+
       0 |      5 |      0 |      5
---------+--------+--------+
       1 |      0 |      5 |      5
---------+--------+--------+
Total           5        5       10

Frequency Missing = 100
```

Output 7. Classification table of new observations

This classification table shows the classification results of only our ten new observations, since YACTUAL is missing for the original observations in data sets BOTH and PROBS. Notice that all ten new observations were correctly classified by the model using 0.25 as the cutoff.

Summary

To properly interpret the results from the LOGISTIC procedure, it is important to pay close attention to response level ordering. Parameter estimates, odd ratios and predicted probabilities may all seem backward if the event of interest is not Ordered Value 1 in the Response Profile of the LOGISTIC output. There are several methods available for adjusting the response level ordering to your needs.

With the response ordering set appropriately, predicted probabilities for existing or new observations may be obtained from the parameter estimates, or more easily, by using the OUTPUT statement. With the CTABLE and PPROB= options, you can select a cutoff value to apply to the predicted probabilities, allowing you to classify the observations and create an observed-by-predicted classification table.

References

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, New York: John Wiley & Sons, Inc.

Freeman, D.H., Jr. (1987), *Applied Categorical Data Analysis*, New York: Marcel Dekker, Inc.

Hosmer, D.W., Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

SAS Institute, Inc. (1990), *SAS/STAT User's Guide, Volume 2, GLM-VARCOMP, Version 6, Fourth Edition,* Cary, NC: SAS Institute, Inc.

SAS Institute, Inc. (1992), *SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07*, Cary, NC: SAS Institute, Inc.