Paper AA08-2013

Improved Interaction Interpretation: Application of the EFFECTPLOT statement and other useful features in PROC LOGISTIC

Robert G. Downer, Grand Valley State University, Allendale, MI

ABSTRACT

The interpretation of fitted logistic regression models for students, collaborators or clients can often present challenges. Explanation of significant interactions among continuous predictors can be particularly awkward. The EFFECTPLOT statement and other features in PROC LOGISTIC of SAS/STAT can be useful aids in meeting these challenges. The CONTOUR and SLICEFIT options of this statement are particularly advantageous for more effective displays. Through logistic modeling of Titanic survival data, this paper also illustrates other ODS graphics and output from models with categorical and continuous predictors. Some basic familiarity with logistic regression is assumed.

INTRODUCTION

For a binary response, a logistic regression model expresses the log odds of presence versus absence p/(1-p) as a linear function of the predictor variables. The logistic regression model for predictors $X_1 \ldots X_k$ is expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_1 + \ldots \beta_k x_k$$

The estimated coefficients $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ can be interpreted on the log-odds or odds scale. Indicator variables are coded for categorical predictors and (in the case of 0,1 predictor coding), exponentiation of the estimated coefficient represents the odds of the response at the given level of the categorical variable versus the baseline category. For continuous predictors, exponentiation of the estimated coefficient $\hat{\beta}_i$ represents the estimated odds of the response for a unit change in the predictor X_i

Although odds and odds ratios are still very common in most applications of logistic regression, interpretations involving the estimated probability of the response are becoming a larger part of model interpretation. The fitted probability $\hat{p}$ can be obtained for each observation and creating a tabular 'profile' of estimated probabilities based on combinations of the predictor variables is still an effective way of expressing a final model for an audience. The ODS graphics displays that are now present in SAS/STAT can now result in quick and easy display of the logistic relationship between a continuous response and the fitted probability at various combinations of categorical variables. The focus of this paper involving interaction interpretation is on the display of the estimated probability rather than odds ratios and output involving significance of model terms is not shown. Significance of interaction terms should be investigated separately. There is general correspondence between graphical display and significance in this paper due to the size of the data set and the nature of the predictors.

INTERACTIONS IN LOGISTIC MODELS AND RELEVANT FEATURES IN PROC LOGISTIC

In a logistic regression model, the two-way interaction between a continuous predictor and categorical predictor is fairly straightforward. If these predictors were age and gender, for example, then we'd fit the following model

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_{gender}(gender) + (age)\beta_{age} + \beta_{age*gender}(age*gender)$$

Significance of the estimated interaction coefficient $\hat{\beta}_{age*gender}$ suggests that the effect of age on the odds of the response depends on gender and there is evidence that the fitted logistic S curves (displaying the estimated

probability as a function of age) will have a different slope. With ods graphics invoked, such a display is produced using the PLOTS=EFFECT option on the PROC LOGISTIC line and will be illustrated in Example 1.

If the interaction of interest involves two continuous predictors (eg. age and weight), the interaction interpretation is not nearly as straightforward but the model takes a similar form:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_{weight}(weight) + (age)\beta_{age} + \beta_{weight*age}(weight * age)$$

Significance of the estimated coefficient $\hat{\beta}_{age*weight}$ suggests that the effect of age depends on weight but the generation of any single logistic curve (displaying the fitted probability as a function of one of the predictors) typically requires fixing the value of the other. Generating estimated curves for specific values of the second continuous predictor (eg. median, quartiles) or a table profile of estimated probabilities has been a traditional way of expressing and interpreting such an interaction. A SCORE statement in PROC LOGISTIC is effective for generating specific estimated probabilities and this syntax is illustrated in Example 2. The EFFECTPLOT statement is a new feature in SAS 9.3 that can generate a series of S-curves at specific values of predictors through the SLICEFIT option (Illustrated in Examples 2 and 3). The contour lines and colors from the new CONTOUR option are a very effective and visually attractive improvement for interpreting such interactions. These features are illustrated and discussed in in Examples 3 and 4.

DATA SET

The Titanic survival data set used in this paper is the titanic3 version found in the Vanderbilt biostatistics data repository and originally constructed by T. Cason at the University of Virginia, with reference to Hind(1999). For more information, see http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3info.txt. The variables being considered in this paper are pclass (passenger ticket class, integer 1-3 with 1 being the first class ticket), survival (the binary response), gender, age and number of siblings/spouses aboard. Using only complete data per individual, there are 1046 observations used in the modeling. Age, the number of parents/children aboard as well as the number of siblings/spouses are being viewed as continuous predictors. Age for infants less than one year is given as a fraction one year.

Some of the notable relative frequencies for this data set include: 40.7 percent of the individuals survived, 75.3 percent of the women survived, there was a 96.2 percent survival rate among women with a first class ticket, a 90.3 survival rate for women over 50, and a 10.9 percent survival rate for men with more than one sibling or spouse aboard.

EXAMPLE 1 (ONE CATEGORICAL PREDICTOR, ONE CONTINUOUS PREDICTOR & INTERACTION)

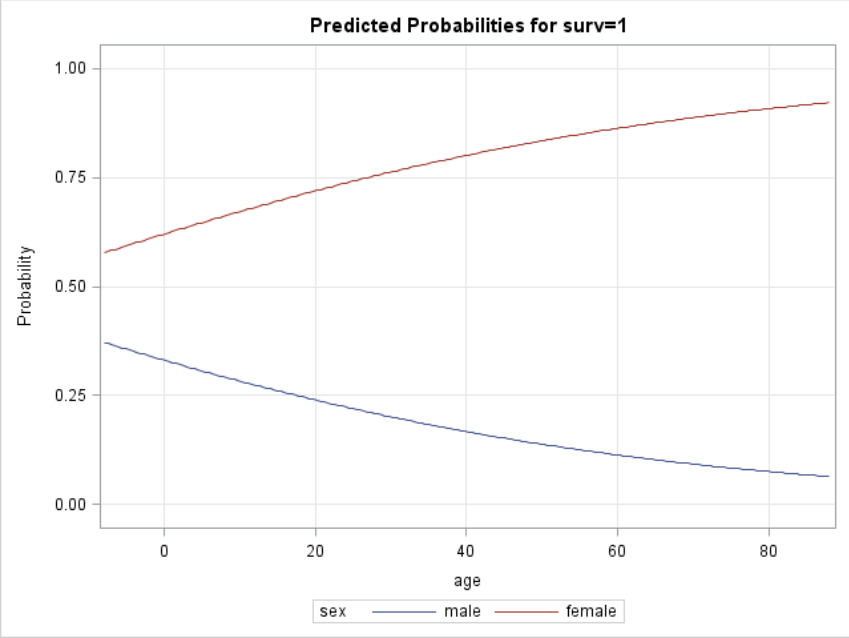
With ODS graphics invoked in SAS/STAT 9.3, consider the following run of PROC LOGISTIC

```
proc logistic descending  plots = effect;
class sex / descending param = glm ;
model surv = age sex age*sex ;
run;
```

The initial 'descending' option on the PROC LOGISTIC line ensures that 1 will be taken as the survival response while the second performs a similar function in the CLASS statement (other options such as 'ref='can also be used for the same purpose). The param = glm on the CLASS statement invokes glm indicator coding as discussed in the introduction.

Interpretation of the interaction displayed through generated logistic curves from the PLOTS=EFFECT option on the PROC LOGISTIC line is straight-forward and is shown in Figure 1. In this example, not only is the interaction significant (suggesting different slopes of logistic curves for each gender) but the direction of the relationship is completely different (output not shown). For female Titanic passengers, the probability of survival increases with age while the probability of survival decreases for men. The default ODS display of the logistic S curves across the levels of the categorical variable is a useful and quick display that is very friendly to the non-statistician and is typically very effective in generating understanding of the differing relationship between the continuous variable and the probability of the response.

Figure 1: Interaction displayed through PLOTS=EFFECT option



EXAMPLE 2 (TWO CONTINUOUS PREDICTORS & INTERACTION)

An initial run of PROC LOGISTIC with predictors age and sibsp (number siblings/spouse) and their interaction shows strong significance of the interaction coefficient ($p < .0001$). An OUTMODEL option on the PROC LOGISTIC line of this model stores the model estimates in the file ex2ests (to subsequently be used below). The significant interaction suggests that effect of age on the odds of survival depends on sibsp and vise-versa. Both are considered continuous variables in this example. Interpretation is not nearly as clear and a beneficial display is not immediate.

Fixing the values of one predictor is one way of approaching the interpretation of the interaction. A separate run of PROC MEANS for each predictor quickly gives ages 21, 39 as the 25th,50th and 95th percentiles of the age distribution and gives 0 and 1 and 2 as the 25th, 50th and 95th percentiles of the sibsp distribution. With missings artificially inserted as the response, a data set PROFILE was created which corresponds to these nine possible 'profile' combinations of the 2 predictor variables.

```
proc logistic inmodel = ex2ests;
score data = profile out = profout;
run;
```

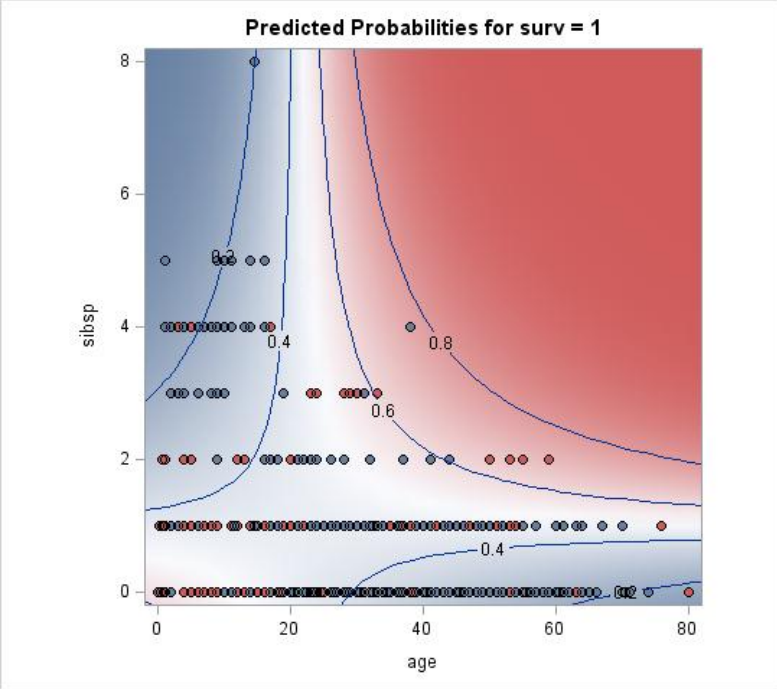
The following Table 1 is the resulting output from applying the SCORE statement to the previously fitted model which has predictors age, sibsp and their interaction. The (default) result variable P_1 is the probability of survival. One can see that with age fixed at 21, the probability of survival is almost constant through the 3 values of sibsp. However at age 39, the estimated probability of survival increases and is at 0.571 for a 39 year old with 2 siblings/spouses aboard. At age 57, the estimated survival probability increases even quicker and is at 0.687 for 2 siblings/spouses aboard. This table is a useful application of the SCORE statement but still has limitations with respect to quick interpretation.

Table 1: Output from LOGISTIC run with SCORE statement

Obs	surv	age	sibsp	P_1
1	.	21	0	0.44707
2	.	21	1	0.44647
3	.	21	2	0.44586
4	.	39	0	0.34669
5	.	39	1	0.45662
6	.	39	2	0.57095
7	.	57	0	0.25832
8	.	57	1	0.46682
9	.	57	2	0.68759

With respect to interpretation, a much more effective output display is given by using the CONTOUR option on the EFFECTPLOT statement given in Figure 2. There is no need to create a file with the estimated probability for specific values of each of the continuous predictors. The contour interpolation shows how the estimated probability changes for different range combinations of the predictors. Default colors and line displays are shown. The characteristics and interpretations of estimated probability contours will be discussed in further detail with the additional predictors and options included in Examples 3 and 4.

Figure 2: EFFECTPLOT's CONTOUR option rather than a 'profile'



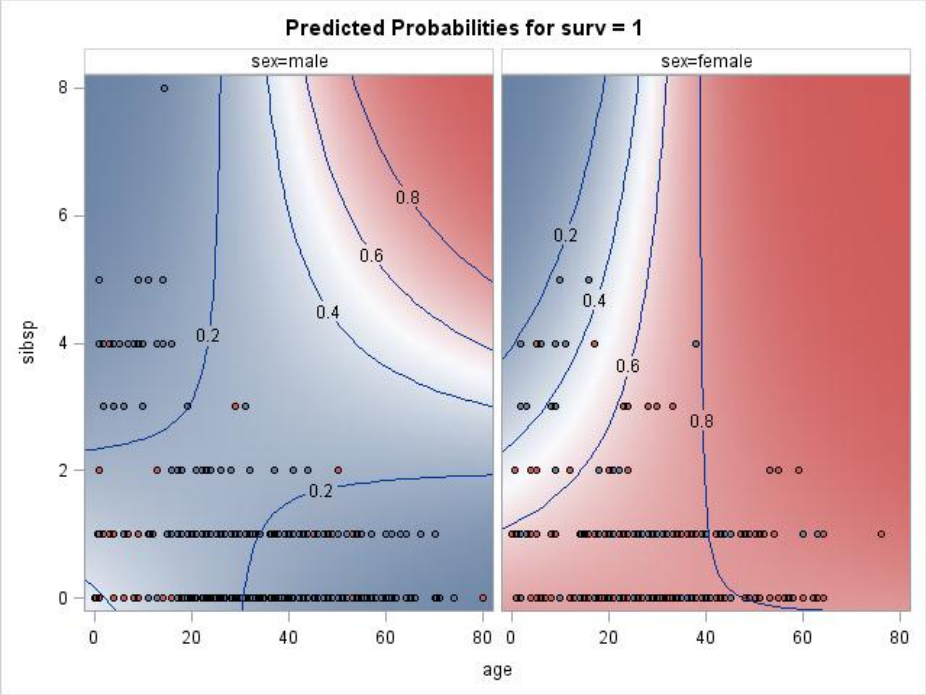
EXAMPLE 3 (ONE CATEGORICAL PREDICTOR, TWO CONTINUOUS PREDICTORS & TWO WAY INTERACTIONS)

The following run of PROC LOGISTIC adds the categorical predictor sex and each of the two-way interactions involving age, sex, sibsp. The (PLOTBY = sex) option is also added and this feature pays immediate dividends here.

```
proc logistic descending;
class sex / descending param = glm ;
model surv = age sex sibsp age*sex age*sibsp sibsp*sex  ;
effectplot contour(plotby=sex) ;
run;
```

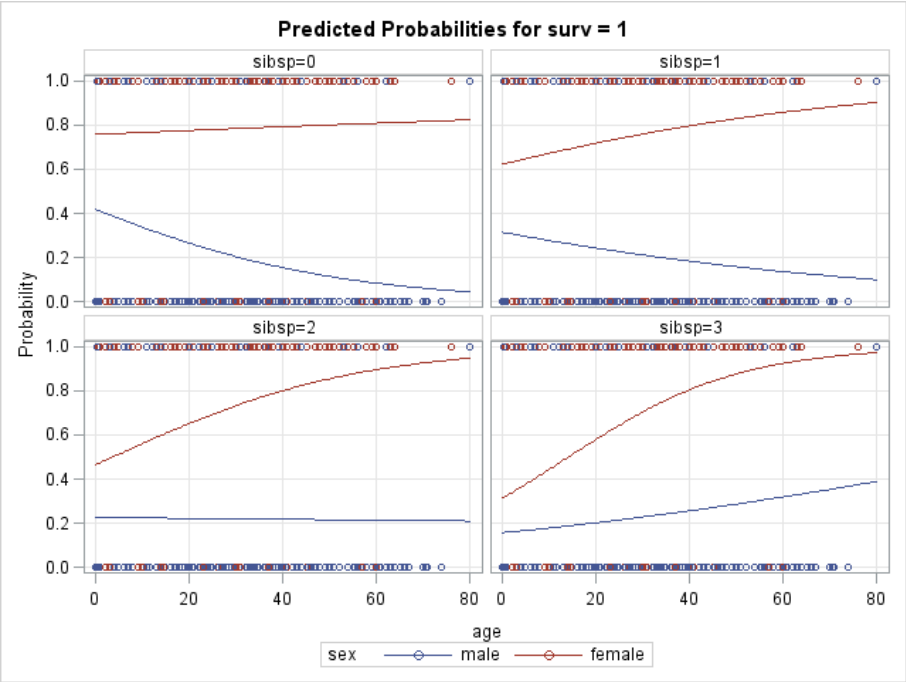
The contour display of Figure 2 is now expanded by gender in Figure 3 and the impact of separating the display by this categorical predictor is quite evident. With survival represented by red points and high estimated survival probability represented by red shading, the higher estimated probability for females is immediately evident. The changing effect of age (depending on gender and sibsp) can be seen within contour plots (for interpretation of the age*sibsp interaction) and across plots (for interpretation of the age*sex interaction). As age increases for females, the estimated probability of survival generally increases but with 4 or more siblings or spouses aboard, a 0.6 probability of survival occurs only after age 30.The pattern is very different for males and the shading is blue (suggesting lower survival probability) for almost the entire plot . This contour plot uses EFFECTPLOT default settings for each of the continuous predictors but much more user input is possible, including axes specifications and control of the observations on the contour plots.

Figure 3: EFFECTPLOT interaction display through PLOTBY option



Alternatively, if one prefers the display of several logistic S curves, one could create a panel of such displays using the SLICEFIT option, perhaps using fixed values of one of the continuous predictors. In Figure 4, we see the actual binary responses (zero or one) well as the estimated logistic S-curves across gender for each of 0,1,2,3 siblings or spouses on board. The direction of the logistic curve (i.e. the sign and magnitude of $\hat{\beta}_{age}$) depends on gender and the value of sibsp, hence giving a quick indication of the significant interaction.

Figure 4: Interaction shown through SLICEFIT option



EXAMPLE 4 (FINAL MODELS & INTERACTION INTERPRETATIONS)

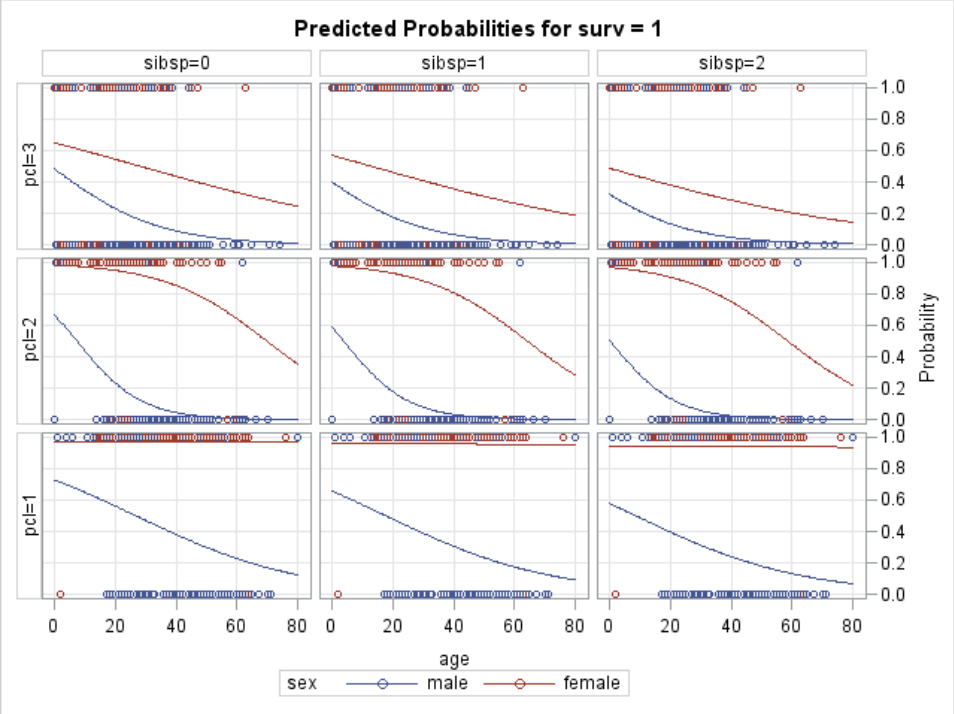
In this Titanic survival data set, the categorical variable pay class (a proxy for socio-economic status) is quite prominent. The survival rates for pay classes 1-3 were 96.2, 89.3 and 47.4 respectively with class 1 being the highest pay class. The survival rate for men in these pay classes was 35.1, 13.9, and 16.9 respectively. The variable pay class pcl (and its two-way interactions with the other predictors) was considered prior to final model selection. Two competing final models had concordance index of 0.862 and 0.864 (deemed not significantly different using ROC concordance area comparison in a separate LOGISTIC run not shown). See Downer and Richardson (2009) for syntax details of ROCCONTRAST.

The following code (using one of the final models) illustrates some of the impressive flexibility of the EFFECTPLOT statement with use of the 'sliceby' 'plotby' and 'at' options. Here, sex, pcl and specific levels of sibsp were chosen for panel categories because age is present in two of the remaining interactions and considering the logistic curves as a function of age allows for intuitive interpretation of the logistic curves.

```
proc logistic data = titan2  descending;
class sex pcl / descending param = glm ;
model surv = age sex pcl sibsp age*sex age*pcl pcl*sex;
effectplot slicefit (sliceby=sex plotby(rows) = pcl) /at(sibsp=0 1 2)  ;
run;
```

In Figure 5, the bottom row of the display clearly shows the high estimated probability of survival for females in the first ticket class (pcl = 1) across each of the levels of sibsp. The estimated probability is near 1.0 for each of the combinations and the interaction of pcl and sex is evident when we note that the logistic curves are quite different for males. The pcl*age interaction is evident differing general shapes of the logistic curves across rows of the display. The significant age*sex interaction is evident throughout each of the individual plots by observing the consistent. differences in logistic curves between males and females within a plot.

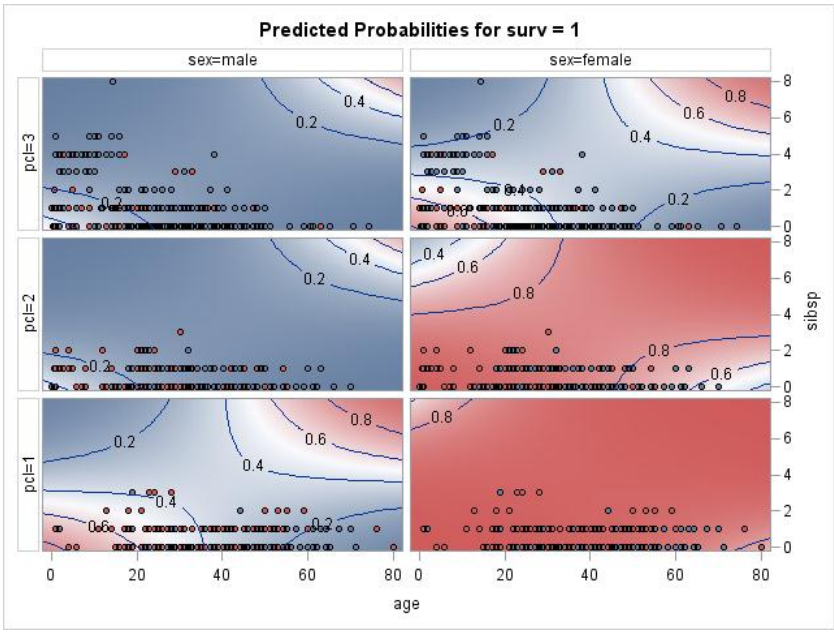
Figure 5: Panel of logistic curves illustrating multiple two-way interactions



The following code uses the other competing model and produces Figure 6. One could argue in favor of this model as the preferred final model based on a parsimony principle since there is one fewer interaction term. Utilizing contours for each of the panel displays (as the possible combinations of age and sibsp) is motivated by presence of the age*sibsp interaction in this model.

```
proc logistic data = titan2  descending;
class sex pcl / descending param = glm ;
model surv = age sex pcl sibsp age*sibsp sex*pcl;
effectplot contour (plotby(rows) = pcl ) /at (sex = 'male' 'female')  ;
run;
```

Figure 6: Contour panel showing interaction across levels of sex & pay class



The effect of age depending on sibsp and the effect of pay class depending on age are very evident in Figure 6. Four of the six contour plots are fairly unique and reflect the significant interactions. A higher siblings/spouse value increased the estimated probability of survival for the oldest passengers. The probability of survival in the lowest socio-economic class (pcl=3) was very low and this was particularly true for males. The red shading in the lower right panels clearly shows that older women in higher pay classes survived the Titanic disaster.

CONCLUSION

New features in PROC LOGISTIC have improved our ability to display and interpret significant interactions within logistic regression models. The capabilities possible through various options of the EFFECTPLOT statement are a substantial improvement and go a long way towards addressing the challenge of interpreting interactions between continuous variables. Visually effective high quality displays result from the default settings but considerable flexibility is possible. Some of these options were reviewed and illustrated in this paper using the Titanic survival data set.

REFERENCES

Downer, R.G. and Richardson, P. R. (2009), 'Illustrative Logistic Regression Examples using PROC LOGISTIC: New Features in SAS/STAT® 9.2 ', Proceedings of the PharmaSUG 2009 Conference, Cary NC: SAS Institute Inc.

Hind, Philip (1999). "Encyclopedia Titanica." Online. Internet. Available http://atschool.eduweb.co.uk/phind

CONTACT INFORMATION:

Robert G. Downer, PhD.
Biostatistics Director & Professor
Department of Statistics, Grand Valley State University
Allendale, MI 49401
downerr@gvsu.edu