Print

Practice: Using PROC LOGISTIC to Perform Backward Elimination; Using PROC PLM to Generate Predictions

The insurance company wants to model the relationship between three of a car's characteristics, weight, size, and region of manufacture, and its safety rating. Run PROC LOGISTIC and use backward elimination. Start with a model using only main effects. The **stat1.safety** data set contains the data about vehicle safety.

1. Use PROC LOGISTIC to fit a multiple logistic regression model with **Unsafe** as the response variable and **Weight**, **Size**, and **Region** as the predictor variables.
    1. Use the EVENT= option to model the probability of Below Average safety scores.
    2. Apply the SIZEFMT. format to the variable **Size**.
    3. Specify **Region** and **Size** as classification variables and use reference cell coding. Specify *Asia* as the reference level for **Region**, and *1* (small cars) as the reference level for **Size**.
    4. Add a UNITS statement with -1 as the unit for **Weight** so that you can see the odds ratio for lighter cars over heavier cars.
    5. Add a STORE statement to save the analysis results as **isSafe**.
    6. Request any relevant plots.
    7. Submit the code and view the results.

```
/*st107s04.sas*/
ods graphics on;
proc logistic data=STAT1.safety plots(only)=(effect oddsratio);
   class Region (param=ref ref='Asia')
             Size (param=ref ref='Small');
   model Unsafe(event='1') = Weight Region Size / clodds=pl selection=backward;
   units Weight = -1;
   store isSafe;
   format Size sizefmt.;
   title 'Logistic Model: Backwards Elimination';
run;
```

Notice that the reference level for **Size** is set to 'Small' in the solution, rather than '1'. When a format is applied to a CLASS statement variable, the reference level option should refer to the formatted value and not the internal value.

Here are the results.

2. Which terms appear in the final model?

Only **Size** appears in the final model.

3. If you compare these results with those from the previous practice (a model fit with only one variable, **Region**), do you think that this is a better model?

Comparing the model fit statistics, you see that the AIC (92.629) and SC (100.322) are both smaller in the model fit by the backward elimination method, 119.854 and 124.982, respectively. This indicates that the **Size**-only model is doing better than the **Region**-only model.

Using the c statistics, you can also see improvement beyond the **Region**-only model, that is, 0.818 in this model compared wtih 0.598 in the previous model.

4. Using the final model that was chosen by backward elimination, and using the STORE statement, generate predictive probabilities for the cars in the following DATA step:

```
data checkSafety;
   length Region $9.;
   input Weight Size Region $ 5-13;
   datalines;
   4 1 N America
   3 1 Asia
   5 3 Asia
   5 2 N America
         ;
run;



 proc plm restore=isSafe;
    score data=checkSafety out=scored_cars / ILINK;
    title 'Safety Predictions using PROC PLM';
 run;

 proc print data=scored_cars;
 run;
```

Here are the results.

Hide Solution