Print

# Specifying a Parameterization Method

PROC LOGISTIC does not work directly with the categorical predictor variables in the CLASS statement. Instead, it first parameterizes, or codes each predictor variable. Just as in linear regression, the CLASS statement creates a set of design variables, also known as dummy variables, that represent the information in each classification variable. PROC LOGISTIC uses the design variables, and not the original variables, in model calculations.

There are various methods of parameterizing the classification variables. Two common parameterization methods, or coding schemes, are effect coding, which is the default in SAS, and reference cell coding. You can choose the reference level in the CLASS statement. It's important to understand a little bit about each method so that you can decide which method to use in your analysis. Different parameterization methods will produce the same results regarding the significance of the categorical predictors, but understanding the parameterization method will help you to interpret your results accurately.

Effect coding is also known as deviation from the mean coding. It compares the effect of each level of the variable to the average effect of all levels. In this example, you can use effect coding to test whether the effect of having a particular income level, such as income level 1, or low income, is different from the average effect of all three income levels (level 2 medium and level 3 high). For effect coding, the number of design variables, or effects, that PROC LOGISTIC creates is the number of levels of the classification variable minus 1. Here, IncLevel has three levels, so PROC LOGISTIC creates two design variables. As the default reference level, PROC LOGISTIC uses the last alphanumeric level. With effect coding, all design variables for the reference level have a value of -1. The design variables for all other levels of the classification variable are set to 0 or 1.

| Value | Label | D1 | D2 |
|---|---|---|---|
| 1 | low income | 1 | 0 |
| 2 | medium income | 0 | 1 |
| 3 | high income | -1 | -1 |

So for example, to estimate the effect of high income on the outcome, the parameter coefficients would both be -1. Whereas if estimating the effect of low income, the coefficient for the first effect would be 1 and 0 for the other.

This formula shows how the parameters are interpreted when effect coding is the parameterization method.

$$\text{logit}\,(p) = \beta_0 + \beta_1 * D_{\text{low income}} + \beta_2 * D_{\text{medium income}}$$

$\beta_0$ is the intercept, but it's not the intercept in terms of where you cross the y axis. Instead, $\beta_0$ is the average value of the logit across all income levels. $\beta_1$ is the difference between the logit for income level 1, or low income, and the average logit across all income levels. $\beta_2$ is the difference between the logit for income level 2, or medium income, and the average logit across all income levels.

Here's the Analysis of Maximum Likelihood Estimates table that PROC LOGISTIC generates for this example.

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -0.5363 | 0.1015 | 27.9143 | <.0001 |
| **IncLevel** | 1 | 1 | -0.2259 | 0.1481 | 2.3247 | 0.1273 |
| **IncLevel** | 2 | 1 | -0.2200 | 0.1447 | 2.3111 | 0.1285 |

Here the parameter estimates and p-values reflect differences from the mean logit value over all levels. The values in the Estimate column are the estimates for the betas defined in the formula. IncLevel 1 designates the design variable for the first level of the IncLevel variable (which is 1, or low income). So, for IncLevel, the Estimate shows the estimated difference in logit values between IncLevel=1 and the average logit across all income levels. The p-value for IncLevel 1 is 0.1273, which indicates that the effect of low income is no different than the average effect of low, medium, and high income. The p-value for IncLevel 2 is also not significant, so the effect of medium income is no different than the average effect of low, medium, and high income.

Reference cell coding compares the effect of each level of the predictor variable to the reference level, which is the last level by default. For example, the effect for the level low income estimates the logit difference between low income and high income. With reference cell coding, all design variables for the reference level have a value of 0. The design variables for all the other levels of the classification variable are set to 0 or 1.

This formula shows how the betas are interpreted when reference cell coding is the parameterization method.

$$\text{logit}\left(p\right) = \beta_0 + \beta_1 * D_{\text{low income}} + \beta_2 * D_{\text{medium income}}$$

$\beta_0$ is the intercept and the value of the logit of the probability when income is high (or at the reference level). $\beta_1$ is the difference between the logit of the probability for low and high income, and $\beta_2$ is the difference between the logit of the probability for medium and high income.

Here's the Analysis of Maximum Likelihood Estimates table using reference cell coding for the classification variable.

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -0.0904 | 0.1608 | 0.3159 | 0.5741 |
| **IncLevel** | 1 | 1 | -.06717 | 0.2465 | 7.4242 | 0.0064 |
| **IncLevel** | 2 | 1 | -0.06659 | 0.2404 | 7.6722 | 0.0056 |

The values in the Estimate column are the estimates for the betas in the formula. For reference cell coding, the meaning of the parameter estimates and p-values is different. Now, the parameter estimate and p-value for IncLevel=1 reflect the difference between IncLevel=1 and IncLevel=3, the reference level.




Close