Binary Logistic Regression Webinar Questions Answered

1. **Is multicollinearity applicable to logistic regression? Can I analyze two correlated (and statistically verified) but significant continuous variables x1 & x2 together?**

   [Smita]: You are correct in expressing concern about the issue of collinearity (essentially, correlation) between your predictor (X) variables. This holds true for binary logistic, OLS regression and indeed for any generalized linear model. If the correlation is small it wouldn't affect your results, but as the two variables are significantly correlated, ignoring this and keeping both correlated variables in the model could lead to a statistically significant overall model, but where each of the individual predictors is insignificant, or have very large coefficients and/or standard errors. For practical purposes, either X1 or X2 would serve equally well as predictors but it is redundant to use both together. I recommend you drop one of the two variables from your model.

2. **On the Late Debt Payment example on slide 22, does it add any value to utilize a "membership" value for the age instead of binary 1 or 0? Would that add any value to the analysis?**

   [Smita]: In general, the higher the measurement scale the better (nominal < ordinal < continuous), except in this case where we suspected that the odds of Default change depending on the age-group. So this is one of the few times that categorizing the continuous predictor is advised- if as I assume, "membership" is a sort of ordinal variable indicating "membership" in the corresponding age group, as opposed to a 'Yes/No' designation, then I don't think it would add any more information to the analysis – but it would likely get similar results. However the ordinal indicator variable might help in case we need to have more than 4 or 5 categories because the odds tend to change over smaller age intervals than those assumed in the current analysis.

3. **How can you identify what a sample size should be based on data with 3 variables? For example, I have data with over 2500 items with 3 variables, Yes, No, N/A, but I want to identify what amount is good sampling to have a 95% confidence level.**

   [Smita]: From the question it appears you have a single variable with three categories. Are you trying to estimate the proportion of 'Yes's? If so, you can simply use the sample size calculation formula for estimation of a proportion:

   $n = Z^2_{\alpha/2} * (p*q/\delta^2)$

   where:

   - $Z_{\alpha/2}$ is the $100(1-\alpha/2)^{th}$ percentile of the standard normal distribution
   - p is the estimated proportion from a pilot sample/historical data
   - q = 1-p
   - δ (delta) is the desired precision of the estimate (e.g. +/- 5%)
   - α (alpha) is the risk/significance level value

If you're using multinomial logistic regression to estimate the factors contributing to the three categories, the sample size would depend on the number of predictors/covariates being considered and the size of the effect to be detected. (*See Note below)

4. **Will you be able to send the way you ordered your data as that is the most difficult part for me to get the data in the proper columns.**

[Smita]: In the 'owncar' example the data were ordered exactly the same way as for linear regression with a continuous response variable – one column lists the individual values (0/1) of the binary Y variable (this is called the 'raw' data) and one column per predictor variable: Income, Age and the binary variable Male (Male=1, Female=0).

Since the last dataset ('default') includes a categorized continuous variable, it was converted to dummy variables based on the age category (<35, 35-64, 65+) each person belonged to. So the two dummy/indicator variables are '<35' and '35-64' with '> 65' serving as the reference group. They're coded as shown below:

	Indicator/Dummy variables for Age	
	< 35	**35 – 64**
Age less than 35 years	1	0
Age between 35 and 64 years	0	1
Age 65 years and older	0	0

Final Note:

I wanted to add a little more detail to my answer to the question during the webcast: **What is a reasonable minimum sample size for odds estimates and how does it compare to OLS estimators?**

Not a whole lot of research has been done on the sample size issue in logistic regression and the few formulas available are quite complicated and are contingent on some rather restrictive assumptions. As I had mentioned one of the things you have to think about is the number of 'cases' or events in the sample – similar to the 'cell counts' in a contingency table.

I want to clarify that it is the number of events per parameter in your model that is at issue. Peduzzi et. al. have shown that the least frequent outcome (whether 'success' or 'failure') in the sample should occur a minimum of 10 times per parameter in order to avoid problems in variance estimation leading to poor coverage of tests and confidence intervals. Of course, as with any other 'rule of thumb', this rule should only be used as a guideline in sample selection.