Cornell University
Cornell Statistical Consulting Unit

StatNews #84

Interpreting Interactions in Logistic Regression

October 2012

Logistic regression is useful when modeling a binary (i.e. two category) response variable. This newsletter focuses on how to interpret an interaction term between a continuous predictor and a categorical predictor in a logistic regression model. We suggest two techniques to aid in interpretation of such interactions: 1) numerical summaries of a series of odds ratios and 2) plotting predicted probabilities. For an introduction to logistic regression or interpreting coefficients of interaction terms in regression, please refer to StatNews #44 and #40, respectively.

To explore this topic we consider data from a study of birth weight in 189 infants and characteristics of their mothers. The response variable is binary, low birth weight status: lowbwt =1 if the birth weight is <2500 grams and lowbwt =0 otherwise. The continuous predictor is the age of the mother in years, and the categorical predictor is whether or not the mother made frequent physician visits during the first trimester of pregnancy, i.e. ftv=0 if no and ftv=1 if yes. To simplify the interpretation of the effect of age by ftv status on the outcome, the age variable was centered at the sample mean of 23 years (i.e., age_c in the model below is equal to age-23). See StatNews #66 for more details on centering. In our model, the log odds of a low birth weight infant is assumed to be a linear function of the two predictors and their interaction:

$$logit(lowbwt) = ln\left(\frac{P(lowbwt = 1)}{1 - P(lowbwt = 1)}\right) = \beta_0 + \beta_1 age\_c + \beta_2 ftv + \beta_3 age\_c * ftv$$

We estimate the coefficients of this logistic regression model using the method of maximum likelihood, with the results reported in the table below.

	β	SE(β)	z	p-value
intercept	-0.52	0.21	-2.44	0.015
age_c	0.04	0.05	0.95	0.344
ftv	-0.47	0.33	-1.41	0.158
age_c*ftv	-0.18	0.07	-2.59	0.010

Note: Dummy coding used where ftv=0 is reference category

Odds Ratios

Although this table tells us we have a significant interaction, interpreting the effect of the interaction term may be challenging. One method to understand the interaction can be through exploring several odds ratios expressing the association between low birth weight and frequent physician visits, at *different* levels of mother's age. The odds ratios in the below table can be calculated using model coefficients reported in the previous table and the following formula:

$$\frac{P(lowbwt = 1)}{1 - P(lowbwt = 1)} = e^{\beta_0 + \beta_1 age\_c + \beta_2 ftv + \beta_3 age\_c * ftv}$$

Recall that an odds ratio of 1 means no association between predictor and outcome (holding other predictors fixed). Odds ratios from the low birth weight example can be summarized as follows.

Mom's Age	OR_{FTV}	p-value	95% Confidence Interval	
17	1.868	0.209	0.705	4.949
23	0.625	0.158	0.325	1.201
24	0.521	0.063	0.262	1.036
25	0.434	0.028	0.206	0.916
30	0.174	0.006	0.050	0.607

For example, the last row shows that a mother at the age of 30 who visits the physician frequently has 0.174 times the odds of having a low birth weight baby as compared to those of the same age who don't visit the doctor frequently, and it is a statistically significant association. For women whose ages are between 17 and 24, the 95% confidence intervals of the odds ratios include the null value of 1 and the interpretation is that the event of having a low birth weight baby is equally likely for the group that frequently visits the physician and the group that does not make frequent visits. However from age 25 and older, the odds of having a low birth weight baby significantly decrease if the mother frequently visits her physician.

Probabilities

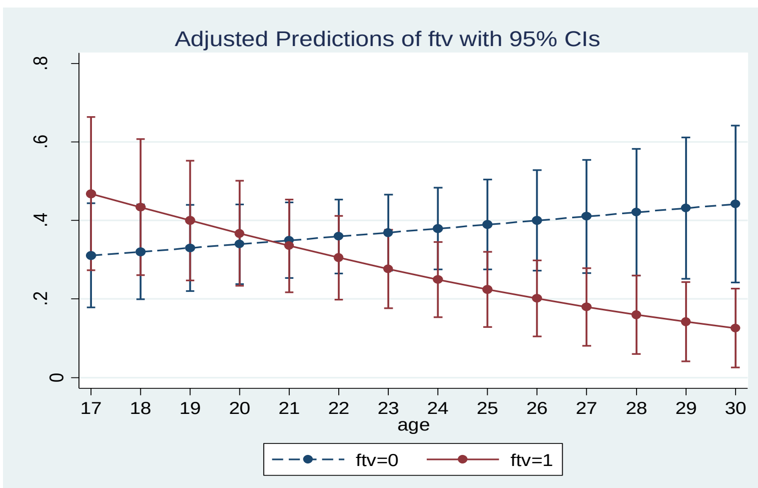
A parallel approach to investigating the nature of this interaction is through calculating predicted probabilities of having a low birth weight infant across different levels of mother's age and frequent physician visits. In this logistic model, predicted probabilities are given by the following equation:

$$\text{P(lowbwt=1)} = \frac{e^{\beta_0 + \beta_1 age\_c + \beta_2 ftv + \beta_3 age\_c * ftv}}{1 + e^{\beta_0 + \beta_1 age\_c + \beta_2 ftv + \beta_3 age\_c * ftv}}$$

Differences in predicted probabilities of low birth weight between those who visit the physician and those who do not (along with p-value for the test if this difference is significantly different from zero) are summarized in the below table for each age group separately.

Mom's Age	Diff. Prob.	p-value	95% Confidence Interval	
17	0.157	0.192	-0.788	0.393
23	-0.092	0.191	-0.197	0.088
24	-0.130	0.072	-0.232	0.046
25	-0.165	0.030	-0.315	-0.016
30	-0.316	0.006	-0.540	-0.092

The results from this method are in agreement with the findings based on odds ratios, although it is noteworthy that the p-values do not have to match exactly between these two metrics. A young mother (i.e. <=24y) has statistically equal probability of having a low birth weight baby whether or not she frequently visits the physician. Then from age 25 and older, there appears to be a significant decrease in the probability of a low birth weight baby if the woman frequently visits the physician. The result of implementing this approach is also summarized in the following plot of predicted probabilities and associated confidence intervals (using Stata 12) and gives a nice visual representation of the interaction.



Importantly, the substantive conclusions that an interaction is present and the direction of the interaction will not be affected by the minor discrepancies that come about from using different metrics. They are equally valid techniques for exploring the nature of an interaction in a logistic regression model.

Both techniques can be implemented in various statistical software packages. If you have any questions regarding implementation or interpretation of these methods, please contact the CSCU Office.

Authors: Hongyu Li and Jay Barry