

HLM

Hierarchical Linear Modeling

Lesson Eight

Logistic Regression

- I. Model and Assumptions
- II. Interpreting the Regression Coefficients
- III. Running Logistic Regressions in SPSS
- IV. Running Logistic Regressions in HLM

Model and Assumptions

Predicting whether an event will occur or not can be of interest in educational research. For example, we may want to model why a student drops out of school (or not), or repeats a grade (or not).

In these models, the dependent variable is dichotomous (it can only take two values). When the dependent variable is dichotomous, two problems arise:

1. The assumptions of the classical linear regression model are violated, e.g. the errors are no longer normally distributed.
2. When using multiple regression, the predicted values are not constrained to fall between 0 and 1, and therefore, they cannot be interpreted as probabilities.

The **logistic regression model** is a multivariate technique that is used for estimating the probability that an event will occur.

Logistic regression is based on estimating the probability that an event will occur. The logistic regression model can be written as:

$$\text{Prob (event)} = e^z / (1 + e^z)$$

Or

$$\text{Prob (event)} = 1 / (1 + e^{-z})$$

Where Z is the linear combination

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

The probability of an event not occurring is:

$$\text{Prob (no event)} = 1 - \text{Prob (event)}$$

When the covariates are centered, the expected probability for an “average” child is simply given by:

$$\text{Prob (event)} = 1 / (1 + e^{-\beta_0})$$

In linear regression, the parameters are estimated using the least squares method. In other words, the reported coefficients are those that correspond to the smallest sum of squared distances between the observed and predicted values of the dependent variable.

In logistic regression, the parameters are estimated using a maximum likelihood technique. In other words, the coefficient estimates are those which make our observed results most likely. The logistic regression model is nonlinear, so parameter estimation uses an iterative algorithm.

Interpreting the Regression Coefficients

In multiple linear regression, the coefficient is the amount of change in the outcome variable for a one-unit change in the independent variable. In logistic regression, however, the interpretation is not as straightforward.

The logistic regression model can be rewritten in terms of the odds of an event occurring. The odds of an event occurring are the ratio of the probability that the event *will* occur to the probability that the event *will not* occur. For example, the odds of rolling a 6 with one die are $(1/6)/(5/6) = 1/5$.

Rewriting the logistic regression in terms of the odds, we get:

$$\text{Prob(event)} / \text{Prob(no event)} = [e^z / (1 + e^z)] / [1 - e^z / (1 + e^z)]$$

$$= e^z = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$$

e raised to the power β_i is the factor by which the odds change when an independent variable increases by one unit. If β_i is positive, the factor will be greater than 1, which means that the odds of the event occurring are increased. If β_i is negative, then this factor is less than 1, so the odds are decreased. When β_i is 0, then the factor is 1, which means that the odds are unchanged.