Diagnostics for Logistic Regression

An important part of model testing is examining your model for indications that statistical assumptions have been violated. This *diagnostic* process involves a considerable amount of judgement call, because there are not typically any definitive statistical tests that can be used to provide assurance that the model meets assumptions or not. One reason that diagnosis is somewhat of a judgement call is that assumptions, such as normality of errors, applies to the population, which we do not have definitive information about. This means that the sample data may be expected to depart from what is expected by the assumption even if there is no real violation in the population.

Assumptions with Logistic Regression

I will give a brief list of assumptions for logistic regression, but bear in mind, for statistical tests generally, assumptions are interrelated to one another (e.g., heteroscedasticity and independence of errors) and different authors word them differently or include slightly different lists. I will not discuss several assumptions—independence of errors/observations, correctly specified model (all relevant predictors included), correct functional form, absence of multicollinearity, fixed predictors (measured without error)—in detail here, because they are common to ordinary least squares regression (see Cohen, Cohen, West, & Aiken, 2003, for a good introduction). There are a couple of other special numerical problems that occur with logistic regression that I will also address here.

An important assumption of logistic regression is that the errors (residuals) of the model are approximately normally distributed. The observed values on the response variable cannot be normally distributed themselves, because Y is binary. But the model has a nonlinear transformation of the predicted values, so the degree to which observed values deviate from the predicted values is expected to vary across a range of values, with most residuals being near 0 and fewer residuals deviating far from the predicted line (either above or below). Strictly speaking, the errors are expected to follow a logistic distribution in the population. With a sufficiently large sample size, the normal distribution can be and is typically used as a comparison, because z and χ^2 distributions can be conveniently used for gauging whether values are extreme or not (though this is not a significance test of the distributional assumption, just a method of examining the degree of departure from the logistic distribution). The error distribution assumption pertains to several potential data problems, including skewness, kurtosis, outliers, and heteroscedasticity (larger residuals for some values of X compared with others). These issues are not independent of one another either. Outliers (extreme values) lead to skewness of the error distribution and kurtosis and skewness are closely related mathematically.

Several authors have pointed out that omitted variables that are related to the outcome can bias logistic regression coefficients for the predictors included in the model even if the omitted variables are unrelated to the predictors, a phenomenon known as unobserved heterogeneity (e.g., Allison, 1999, Hauck et al., 1991; Mood, 2010). The impact of omitted variables in logistic regression is in contrast to what occurs with ordinary least squares regression, in which omitted variables have no impact on model coefficients if they are unrelated to the predictor. Unobserved heterogeneity leads to logistic coefficients for predictors in the model that are biased toward zero (i.e., whenever variables are omitted from the model the effects of the variables will be underestimated). The unobserved heterogeneity bias increases for omitted variables that are more strongly related to the outcome and when omitted variables have larger variances (Mood, 2010). Unobserved heterogeneity complicates comparison of odds ratios across samples, across groups, time points, or across different scales because of the sensitivity of odds ratios to predictor scaling and unobserved heterogeneity. Buis (2015) argues that the unobserved heterogeneity phenomenon is a natural consequence of predicting probabilities, because including any variables that account for variance in the outcome, even if unrelated to other predictors, implies that the predicted probability of event occurrence is farther from chance (i.e., $\hat{\pi} = .5$).

Diagnostics

Let's start with a discussion of outliers. In ordinary least squares regression, we can have outliers on the X variable or the Y variable. With logistic regression, we cannot have extreme values on Y, because

observed values can only be 0 and 1. For identifying problematic cases, we therefore need to consider the residuals rather than the observed values of Y. In logistic regression, the residual is defined as the difference between the observed probability that Y = 1 compared with the predicted value that Y = 1 for any value on X. We will use a subscript j to indicate a particular case or group of cases with the same value on X, so the observed probability for some particular value of X is $P\left(Y_j = 1\right)$ and the predicted probability for some particular value of X is $\hat{P}\left(Y_j = 1\right) = \hat{\pi}_j$.[1] The residuals are typically given in terms of frequencies, so the count of the observed values where Y_j = 1, we will call y_j and the count of cases predicted to be to be 1 is $n_j\hat{\pi}_j$, with n_j representing the number of cases with the value X_j. The raw residual, then, is simply the deviation between the observed and expected counts for Y_j = 1, given as $y_j - n_j\hat{\pi}_j$. The *Pearson*, or sometimes *standardized residual*, divides by the standard error estimate (with the number of cases with value X_j given as n_j) is

$$\text{Residual}_j = r_j = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}}$$

It may be a little difficult to imagine the predicted values for Y_j if you think about individual cases with a unique X_j value, but recall that the predicted value is a theoretical value represented by the line that summarizes the X-Y relationships. These values can be evaluated in terms of a normal distribution and the sum of their squared values is often used as chi-squared value representing the overall degree to which the residuals deviate from the line, $\chi^2 = \sum r_j^2$. Alternatively, the deviance residual is sometimes used (corresponds to the studentized residual in OLS), but it is based on G^2 log function, so a bit more complicated

$$Y_j = \begin{cases} 0 & d_j = -\sqrt{2\left|\ln\left(1-n_j\hat{\pi}_j\right)\right|} \\ 1 & d_j \pm\sqrt{2\left|\ln\left(1-n_j\hat{\pi}_j\right)\right|} \end{cases}$$

In multiple logistic regression, we have to consider multiple X values, and so texts often consider a *covariate pattern* using vector notation to refer to a particular constellation of values on a set of predictors, $\boldsymbol{x}_j$ instead of casewise values of X with one case per value.

Because observed values on Y cannot be outliers themselves, there is a considerable focus on identifying potentially extreme values on X. Moreover, with logistic regression, the residuals are dependent on value of X. A common diagnostic index for extreme values on X is *leverage*, or sometimes "hat" values, denoted h_j here.

$$h_j = \left[n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)\right]\left(b_j\right)$$

where b_j is a multivariate measure of weighted distance from the central mean.[2] In ordinary least squares, higher values on h_j reflect more extreme values on X. But Hosmer and Lemeshow (2000) demonstrate that leverage drops off precipitously for very high or very low expected probabilities, so it is problematic as an outlier index. b_j is a better diagnostic for outliers on X then. Leverage values are important though, because Pearson residuals are a direct function of these values

---

[2] $b_j = \mathbf{x}_j'\left(\mathbf{X}'\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{x}_j$, which is closely related to Mahalanobis distance and is not to be confused with the regression coefficient.

$$r_j = \left(y_j - n_j\hat{\pi}_j\right) / \left(1 - h_j\right).$$

Finally, influential cases can be identified by exploring the degree to which the model fit or the coefficients are altered by removing a particular case. $\Delta\chi_j^2$ is the change in the model chi-square by deletion of a single case (analogous to standardized deleted residuals), ΔD_j is the change in the deviance by deletion of a single case (analogous studentized deleted residuals), and $\Delta\beta_j$ is the change in the regression coefficient by deleting a case, known as ***dfbeta***. Each of these indices have a value for each case in the data set. The fit or coefficient for the model is computed repeatedly deleting one case each time using all cases except the j^{th} case. Authors sometimes recommend cutoff values for these indices, but it is best to obtain the values for all cases and investigate cases for which the value is high relative to other cases in the data set.

Visualization

A critical step in evaluating model assumptions should be plots of the data. We can use any of these various diagnostic values in a plot, usually putting the estimated probabilities, $\hat{\pi}_j$, on the x-axis. The estimated probabilities (i.e., analogous to predicted values in OLS) stand in for X values in a multiple regression, because they are a perfect weighted function of the set of predictors in the model. Menard (2010) and Hosmer, Lemeshow, and Sturdivant (2010) illustrate several types of plots, and I show how to obtain a couple of them below.

Remedies

There are several potential problems outlined above, and there are remedies for most these issues, although not all ideal. For outliers, there are several options, including identifying an entry or computational error and correcting it, eliminating an invalid case (e.g., did not meet inclusion criteria), transforming the relevant variable, analyzing the data with and without the outlier and reporting both sets of results, or use an alternative estimation or robust approach. Dependence of observations (errors) implies some type of clustering in many instances, which may result from nesting (e.g., within household or organization) or serial dependency or time-related clustering (e.g., longitudinal data). Dependence may be addressed with a robust estimator or explicit modeling of clustering. Robust estimators (e.g., Huber-White estimates, Huber, 1967; White, 1980; M-estimates, Huber, 1964) may be the most relevant when there is not a design-related complete clustering in which cases are nested within organizations or observations are nested within individuals (i.e., longitudinal data). For design-related clustering, complex sampling design adjustments (e.g., see Lee & Forthofer, 2006), generalized estimating equations (GEE; Liang & Zeger, 1986) or multilevel regression models (aka hierarchical linear models; Raudenbush & Bryk, 2002) can be used.

.