The Hosmer-Lemeshow Goodness of Fit Test: Does the Grouping Really Matter?

Rivera H, Williams DK, Bursac Z, Hosmer D

Introduction

When considering a predictive model, it is important to evaluate the goodness of fit. Goodness of fit refers to how well the independent variables in a model predict the outcome. More specifically, it indicates how far the actual data deviates from the prediction. This measure is particularly valuable in logistic regression when a set of independent variables is used to predict a binary outcome. Logistic regression is widely used in public health studies. Available in most software packages, it is easy to estimate a subject's probability of the outcome and determine odd ratios from the estimated coefficients. (Hosmer, Lemeshow, & Taber, 1991) It is important to test the fit of a model using a reliable and powerful method in order to draw correct inferences from the model.

The Hosmer-Lemeshow goodness of fit (HLGoF) statistic is one such statistic that is useful in assessing the quality of a model's fit. It is calculated by grouping the predicted probabilities into deciles then examining the difference between the observed and expected frequencies of the outcome. (Shah & Barnwell, 2003) There has been some question regarding the statistical accuracy of the grouping method used to calculate the statistic. Recent speculation has implied an alternative grouping method may yield a different result. This study proposes calculation of the statistic using an alternative method that will sort the predicted probabilities into ten groups at random. By conducting several simulations, we will fit a logistic model and alter some parameters that deviate the fit. This, in turn, can be used to output a goodness of fit statistic calculated using both methods. The goal is to examine the similarities or differences between the resulting HLGoF statistics while testing the performance of both methods to detect departures from the prediction equation.

Background

Building a good model is essential before assessing goodness of fit measures. A model may be thought of in two principle components: a systematic and error component. (Hosmer, Lemeshow, & Taber, 1991) The systematic component is *y* as a function of the independent variables in the model. The values that comprise this group of covariates yield an associated *y* value. The error component is representative of the difference between the observed value of *y* and the expected value given by the values of the covariates. In public health research, models are created using variables of both biological and statistical importance. (Hosmer, Lemeshow, & Taber, 1991) Although it is sometimes necessary to include a variable not necessarily significant to control for confounding. (Hosmer, Lemeshow, & Taber, 1991) One issue that may occur when constructing a model is the covariate pattern. This term refers to the set of values for the covariates in the model. Goodness of fit is evaluated by the fitted values of these patterns. The difference between the observed and fitted values will be referred to as the summary measures.

The summary measures can be thought of as the difference between the *y*-fitted and *y*-observed. A small value does not necessarily indicate a gross lack of fit; however, exceedingly large values indicate a fundamental error in the model's construction. (Hosmer, 2007) In logistic regression, each covariate pattern has some fitted y-value associated with it and an estimated probability. The residuals give way to calculating the

Pearson Chi-Square statistic. The disadvantage with this method is that when the covariate pattern increases as the sample size increases, therefore so do the number of parameters. In this case, the calculated p-values may be incorrect. It is for this reason that Hosmer proposed grouping by deciles. This method is preferred over the fixed grouping method because it follows a $X^2_{df=8}$ distribution more closely.

Regression diagnostics are often employed to ensure the residual variation in a model is small. The Hosmer-Lemeshow decile grouping method is unique in that it examines the differences of observed and expected values within these groups based on the estimated probabilities. Despite its utility, the HLGoF statistic is prone to miss deviation from perfect fit due to a small number of datapoints. (Hosmer & Lemeshow, 1989) A study conducted by Hosmer, Hosmer, Le Cessie, & Lemeshow (1997) on various goodness of fit measures asserted that since the grouping is based on estimated probabilities, *y* as a function of *x*, it may lack power to detect departures in *x*.

This same study evaluated the Hosmer-Lemeshow decile grouping method with the method using fixed cutoff points. The two methods were used to evaluate the same fitted model at a sample size of 100, the two statistics produced widely different values. However as the sample size increased to 500, the two statistics produced nearly the same results. The power to detect deviations increased with sample size.

The fundamental test of the Hosmer-Lemeshow statistic is the null hypothesis that the model fits the data, and the alternative is the model does not fit the data. The statistic for this test is given by examining the differences between the observed and expected frequencies within the deciles where the observed and expected frequencies can be given by the set of equations below.

$$O_{1j} = \sum_{i \in D} y_i \qquad O_{0j} = \sum_{i \in D} (1 - y_i)$$

$$E_{1j} = \sum_{i \in D_j} \hat{\Pi}_j \qquad E_{0j} = \sum_{i \in D} (1 - \hat{\Pi}_j)$$

The summation of the differences across the deciles yields a single summary measure. That can be given by this equation:

$$\hat{C} = \sum_{k=0}^{1} \sum_{j=1}^{10} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}$$

In which, the value of C is expected to be small if the differences between the observed and fitted values are small. Hosmer, Hosmer, Le Cessie, & Lemeshow (1997) noted the variation between HLGoF statistics produced in different software packages. The differences were based on variations in the algorithms used to define the groups suggesting that the statistic may be sensitive to the groupings. Since it has a X^2 distribution, it is also sensitive to small expected values. Thus a visual inspection of the deciles, in combination with other regression diagnostic tests, is the best way to fully evaluate the fit of a model.

Methods

For this study, we used a statistical software package, R, to conduct a simulation study that would allow us to calculate the HLGoF statistic originally as Hosmer proposed and again using our proposed alternative random grouping method (HLRG). First we needed to construct a model where the fit would be nearly perfect and then alter the parameters of that model to alter the fit. We used R to fit a model, where the logit=β_0+0.08x+β_2 x^2, and where x was a uniformly distributed continuous value with a range from -6 to 6. Therefore, each value in the range had an equal chance of being selected. The model was fit where the true model under the null was $G(x) = \beta_0+0.08x$. This model represents nearly perfect fit. Altering the value of the quadratic term allowed us to change the fit and consequently test to see if the goodness of fit statistic would detect these deviations. The model allowed for simulations with various sample sizes and values of β_2 (non-centrality index) coefficients.

First we performed several simulations to fit the null based on 5 different sample sizes 50, 100, 200, 400, and 800. At each level of sample size, we increased the value of the non-centrality index in order to introduce some deviation from the null. The non-centrality index was increased from 0 to 0.08 by increments of 0.01 units. In this way, we can identify the proportion of times the HLGoF test rejects as the fit becomes worse. The more the true logit conforms to a parabola, the greater the proportion of rejection. These combinations yielded 30 unique simulated datasets comprised of the HLGoF X^2 values, their respective *p*-values, HLRG X^2 values, and their respective *p*-values. To visualize the data, we used *ggplot2*, a package in R, to create figures that would allow us to visually inspect the results of the simulation.

Results

To compare the performance of both methods, we used R to plot a facet graph of the HL X^2 values along the y-axis and the HLRG X^2 values along the *x*-axis. Each facet represents a different simulation scenario. The X^2 values produced for both methods in each simulation scenario are plotted against each one-to-one. The blue line is a loess model fit of the values. If our hypothesis is true, that there is no difference in the statistics calculated by the two methods, then you should expect to see the loess trend line follow the small grey line, meaning both methods are producing similar values. With the exception of the scenario where the sample size equals 400 and the non-centrality index equals 0.00, the blue loess line favors the X^2 values generated using the decile grouping method. That is, for larger X^2 values produced by the original method, there are smaller X^2 values being produced by the alternative method (Figure 1).

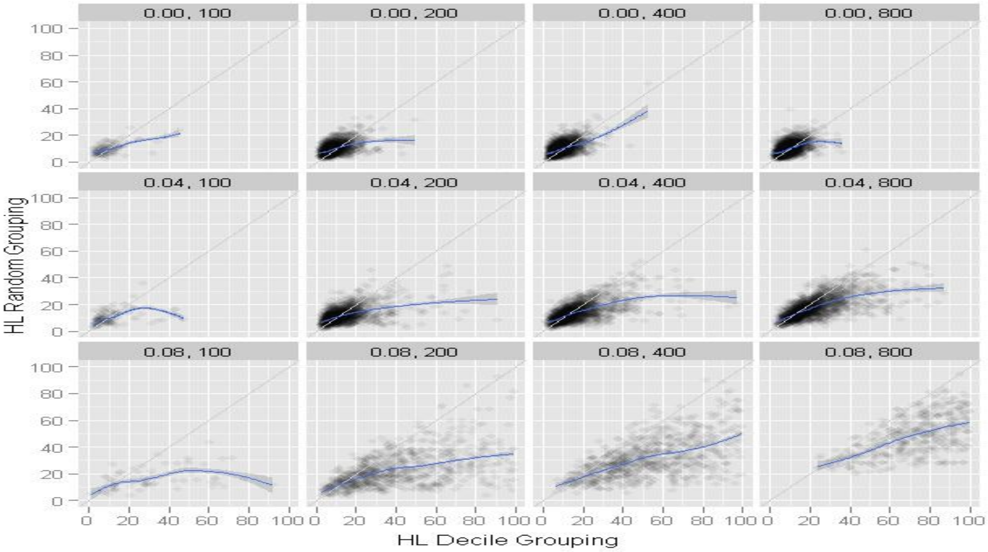

Figure 1 shows the X^2 values plotted against each other calculated using both methods for various scenarios.

Next we examine how the tests agree for each sample size as the deviation from the null increases. This comparison illustrates the proportion of times the two tests make the same decision. That is, the two statistics agree whether to reject or not reject the null, regardless if that decision is correct. For all the sample sizes where the non-centrality index is equal to 0, the proportion of agreement is around 0.85. Stated another way, the two methods have the same conclusion about 85% of the time. At a non-centrality index of 0.01, the two tests agree for all sample sizes. For a non-centrality index of 0.02, the agreement starts to vary by sample size with the tests agreeing the least at 0.80 for a sample size of 800. The two methods agree the most at 0.85 for a sample size of 100. As the non-centrality index increases, the agreement between the two methods varies significantly by sample sizes. This can especially be observed when the non-centrality index equals 0.08. Here for a sample size of 800, the two methods agree 100% of the time. When the sample size is 100, the agreement between the two tests is 0.60 (Figure 2).

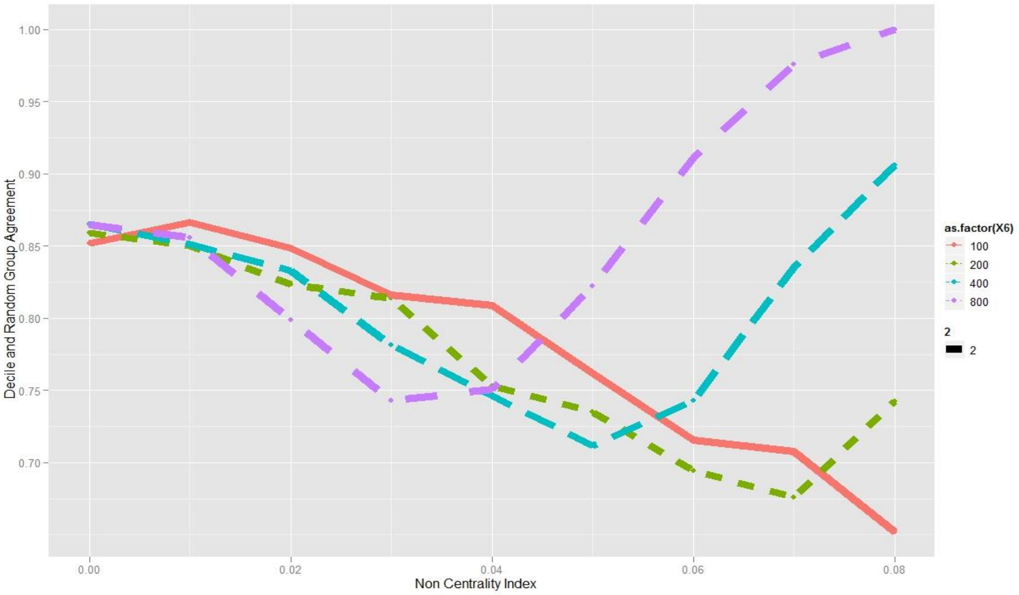

Figure 2 shows the proportion of agreement between both tests.

To compare the performance of the two methods, we constructed two graphs that would allow us to see the proportion of rejection for each method at each sample size and each level of the non –centrality index. Figure 3 depicts the proportion of rejection for those X^2 values generated by the original grouping method. At each sample size, when the non-centrality index is equal to 0, the proportion of rejection is around 0.10. For a 0.05 significance level test, we expect 5% rejection. Therefore our type I error should be around 0.05. With this method, the test rejects the null hypothesis that the model fits the data well about 10% of the time when the model is perfect. Therefore, the empirical type I error rate for this test is about 10%. The proportion of rejection stays fairly similar for all sample sizes when the non-centrality index equals 0.01. At a non-centrality index of 0.02 the proportion of rejection begins to increase rapidly for larger sample sizes as compared to smaller sample sizes. For a sample size of 800, the proportion of rejection starts to plateau as it reaches 1.00. The other samples follow the same pattern but at a much slower rate. When the non-centrality index is equal to 0.08, the proportion of rejection is 1.00.

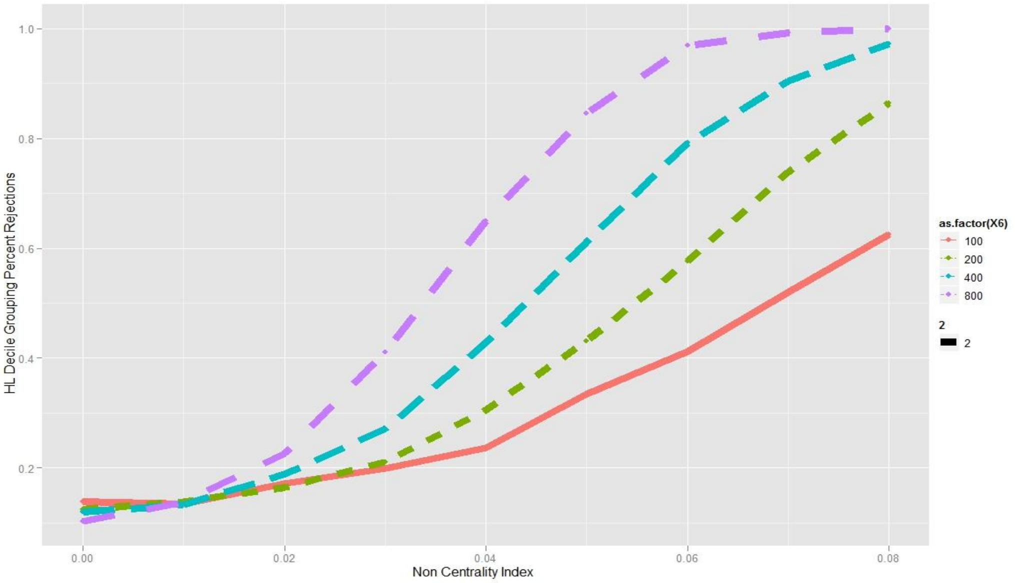

Figure 3 shows the proportion of rejection for the X^2 calculated using the original method of calculation.

Figure 4 depicts the proportion of rejection for the HLRG method. When the non-centrality index is equal to 0, for each sample size the proportion of rejection is around 0.10, much like the original grouping method. The greater the sample size the greater rate of increase in the proportion of rejection. Therefore, at a non-centrality index of 0.06 for sample size of 800, the proportion of rejection is 0.98 and is subsequently smaller for each decrease in sample size. At a non-centrality index of 0.08, when the sample size 800 the proportion of rejection is the highest at 1.00 and is the lowest when the sample size is 100 at 0.38. The contrast between the two graphs shows that the original grouping method rejects at a higher proportion and a faster rate when the non-centrality index and sample size increases.

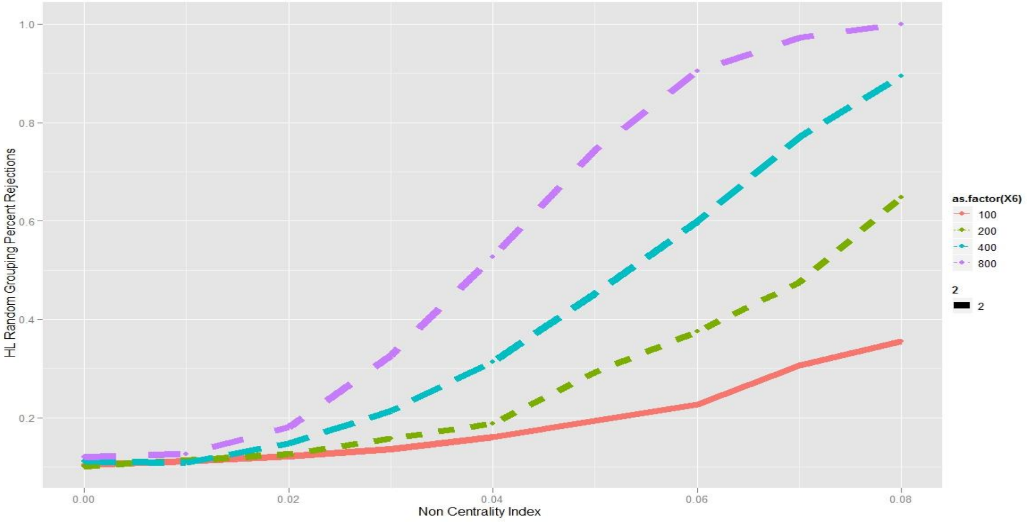

Figure 4 shows the proportion of rejection for the X^2 calculated using the alternative method of calculation.

Discussion

Upon inspection of our graphs, you see both statistics are very aggressive in evaluating the model. When the model is nearly perfect, both tests reject about 10%. For a 0.05 significance level test, we expect the proportion of rejection to be around 0.05. However, this was not observed and therefore our observed type I error rate is approximately 10%. As the non-centrality index increases so does the proportion of rejection, as expected. However the original grouping method is more sensitive to this increase, and the proportion of rejection increases at a faster rate as compared to the alternative method. In addition the power to detect deviations from the null increases as the sample size increases.

The alternative method is more conservative than the original method. A comparison of both methods, where the points where the sample size is equal to 100 and the non-centrality index equals 0.08, shows that the original method rejects 62% and the alternative method rejects 38%. Considering these two points gives us some insight to the decline in agreement between the two tests as shown by graph 2. Shown in graph 1, the original method is producing larger values and is therefore rejecting the goodness of fit test more often. Although the two tests are more aggressive when the model is nearly perfect, the original grouping method is far more sensitive and powerful to detect departures from the null. Therefore we can conclude that the original grouping method is far more superior to detecting deviations from good fit.

Works Cited

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics Review 14: Logistic Regression. *Critical Care , 9* (1), 112-118.

Hosmer, D. (2004). Assessing the Fit of Logistic Regression Models. In D. Hosmer.

Hosmer, D., & Lemeshow, S. (1989). *Applied Logistic Regression* (2nd ed.). New York, NY: John Wiley & Sons, INC.

Hosmer, D., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model. *Statistics in Medicine , 16*, 965-980.

Hosmer, D., Lemeshow, S., & Taber, S. (1991). The Importance of Assessing the Fit of Logistic Regression Models: A Case Study. *American Journal of Public Health , 81* (12), 1630-1635.

Sarkar, S., & Midi, H. (2010). Importance of Assessing the Model Adequacy of Binary Logistic Regression. *Journal of Applied Sciences* , 479-486.

Shah, B., & Barnwell, B. (2003). Hosmer-Lemeshow goodness of fit test for Survey Data. *2003 Joint Statistical Meetings-Section on Survey Research Methods* , 3778-3780.