Examining Model Fit for Logistic Regression on Large Data Sets

Todd Connelly * Trent L. Lalonde, Ph.D. †

Abstract

The Hosmer Lemeshow Test (HLT) is commonly used as a goodness of fit test for logistic regression. However, it is over-powered in medium (100,000 to 500,000 observations) to large (1 million plus observations) datasets. Recent research [Paul, Pennell, Lemeshow 2012] proposes to address this by increasing the number of groups for the HLT to disperse the power. This helps expand the HLT to datasets of up to 25,000 observations. Yet, in today's world of big data we need to be able to assess fit on logistic regression models with large datasets. We propose a bootstrapping approach to obtain a modified HLT (mHLT) statistic. Several point estimates are considered for being the mHLT, including a median, trimmed mean and 5th and 95th percentiles.

Key Words: Big Data, Goodness of Fit, Logistic Regression

1. Introduction

Logistic regression is a commonly used statistical model when you have two possible outcomes. It has been used to predict if a customer will purchase again (yes or no), mortality outcomes (alive or dead) and many more. One of the most common tests to assess if your model fits your data well, is called the Hosmer-Lemeshow Test (HLT). This test has been used for many years and is common in modern statistical software packages.

After reading a paper by Lemeshow [7] the issue became clear. The power of the HLT increase very quickly with sample size. This is true with most statistical tests however, it is not a desirable property with a goodness of fit test. The issue of power increasing rapidly with sample size was also[1] found by Kramer and Zimmerman [5] while studying mortality.

1.1 Current State

Currently, there are 3 different data situations that exist for applying the HLT to determine model fit. The first is when your sample is small ($n \leq 1{,}000$) then setting your groups equal to ten. The second case is where your data is slightly bigger ($1{,}000 \leq n \leq 25{,}000$) and you have to use an equation to determine the number of groups. The third case, which is the most concerning involves larger datasets over 25,000 observations. The HLT is simply not recommended in these situations. This is the area that needs improvement so practitioners can make informed decisions regarding model fit.

2. Potential Solutions

Below are multiple methods that attempt to overcome this limitation and provide a method for determining model fit for large datasets using logistic regression. The goal is to find a

*University of Northern Colorado

†University of Northern Colorado

[1] I encountered this issue while running a large logistic regression (over a million observations), at work. After reconsidering the model I had specified and trying several other models that made sense, I found myself still unable to 'fail to reject' on a HLT.

way to modify the HLT or potentially apply it more than once to different parts of the data in order to have a valid test statistic with larger datasets.

Proposed Solutions:

1. Split the data set
   (a) Fix the number of successes in each sample
   (b) Divide the data set into a number of small dataset where the HLT is known to work using SRS.
2. Subsampling
   (a) Use small samples and test the HLT repeatedly with the group size fixed at ten (g=10).
   (b) Use larger samples of 25,000 and test the HLT repeatedly with the group size fixed at some appropriate yet undetermined level.

One thing that is solved by the approach in 1a is that the number of groups 'g' would be constant for each sample. The samples could simply take all of the successes and split them into 'q' mutually exclusive groups. Simple Random Sampling (SRS) could then be performed to on the remaining observations (failures) to complete each sample.

If SRS alone was performed 1b there could be a different number of success in each sample which might lead to a different number of groups, see Table 1 (Each group size is in bold that would be selected by using Equation 1) . Then one is left with the problem of trying to draw a meaningful conclusion from a collection of HLT statistics that were performed with varying number of groups. Equation 1 is reproduced from [7] showing how to calculate the number of groups for samples larger than 1,000 observations.

$$g = \max\left(10, \min\left[\frac{m}{2}, \frac{n-m}{2}, 2 + 8\left(\frac{n}{1000}\right)^2\right]\right) \tag{1}$$

Where:
m = number of success
n = sample size

It was determined that approach 1b was a dead end and our time was invested in a repeated sampling method.

Sample	n	Success	Failures	$\frac{m}{2}$	$\frac{n-m}{2}$	$2 + 8\left(\frac{n}{1000}\right)^2$	df
1	10,000	4,000	6,000	2,000	3,000	**802**	800
2	10,000	1,000	9,000	**500**	4,500	802	498
3	10,000	500	9,500	**250**	4,750	802	248

Table 1: Example of Difficulties of using SRS on group size

It is worth noting that in the hypothetical example in Table 1, the degrees of freedom are 3.2 times as large if you compare sample 1 to sample 3. This shows why this approach was not pursued.

Variable	Type	Description
PurchaseFlag	Dichotomous (1= Success and 0 = Failure)	1 or 0 flag indicating if a customer made a purchase in 14months following their first order
Order1items	Integer	Values are 1+
ModClassOrder	Factor	The modularity class that a customer is assigned to after their first purchase
Order1Channel	Factor	The Channel that was used to place first order, e.g. Email or Catalog
ordermonth	Factor	Month Factor to help account for seasonality differences that may exist
Order1Sales	Continuous	The amount of money spent on first purchase in US dollars , not adjusted for inflation
EmailStatus	Factor	Email Status at time of data pull, e.g. Opted In or Opted Out

Table 2: Description of Variables

3. Using data

Using our dataset containing 1,118,829 rows we took a 1,000 random samples of size n=1,000 and n=25,000. For each sample a logistic regression model was fit and then a Hosmer-Lemeshow Goodness of Fit Test was performed. The sampling was conducted with replacement. The model that was fitted can be seen in equation 3 . For each pass of the simulation 1,000 observations were taken at random and the model in 3 was fit. Then a HLT test was fit with the number of groups equal to ten. The HLT statistic and the corresponding p-value were saved. This was done repeatedly and generated a distribution of HLT statistics and p-values. These can be seen in figure 4 and 4 . We expect about 5% of the tests to be rejected based on an $\alpha = .05$ assuming that we have the correct model specified. A rejection rate of 0.06 was found which is quite close and may move even close to 5% if the number of iterations was increased.

PurchaseFlag = Order1items + ModClassOrder1 + Order1Channel + factor(ordermonth) + Order1Sales + EmailStatus

(2)

4. Simulation with Sample Size of 1,000

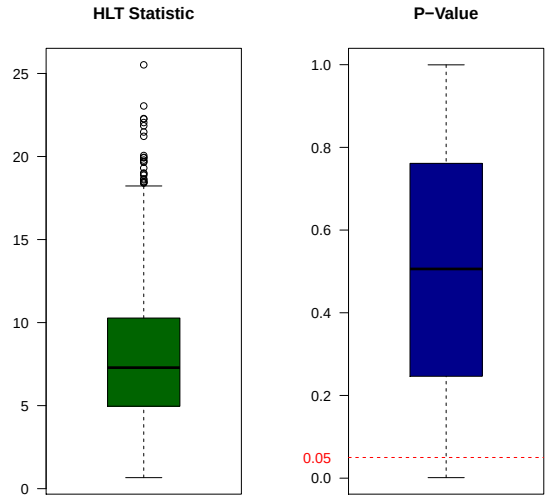

Figure 1: Hosmer-Lemeshow Test Statistic and associated p-values from simulation using n=1,000 with repetitions =1,000

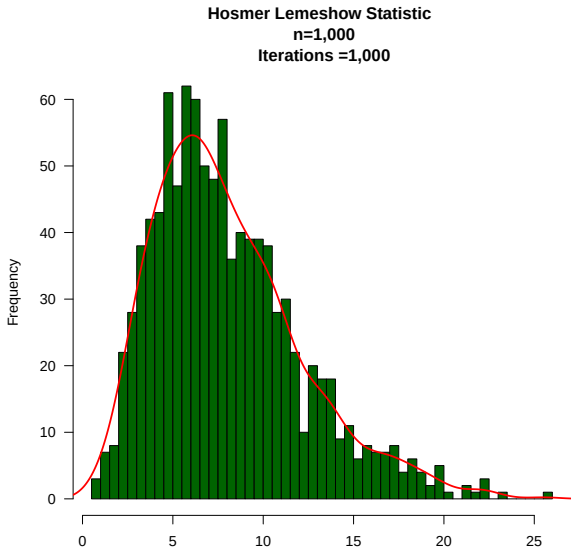

Figure 2: Distribution of the Hosmer-Lemeshow Test Statistic from simulation using n=1,000 with repetitions =1,000

5. Simulation with Sample Size of 25,000

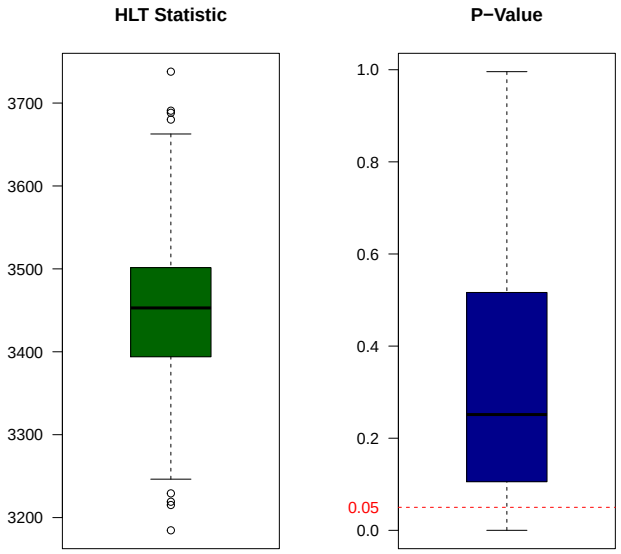

Figure 3: Hosmer-Lemeshow Test Statistic and associated p-values from simulation using n=25,000 with repetitions =1,000

Which statistic would be appropriate to use for determining goodness of fit is not yet clear. Examining the distributions of p-values for both simulation conditions both Tables 3 and 4 are very close and give values that you would expect from a uniform distribution.

Min.	1st Qu.	5th Percentile	Median	Mean	TrimMean	3rd Qu.	95th Percentile	Max.
0.001	0.246	0.040	0.506	0.501	0.502	0.761	0.953	1.000

Table 3: Summary Of P-Values n = 1,000

Min.	1st Qu.	5th Percentile	Median	Mean	TrimMean	3rd Qu.	95th Percentile	Max.
0.000	0.106	0.014	0.251	0.324	0.309	0.516	0.835	0.996

Table 4: Summary Of P-Values n=25,000

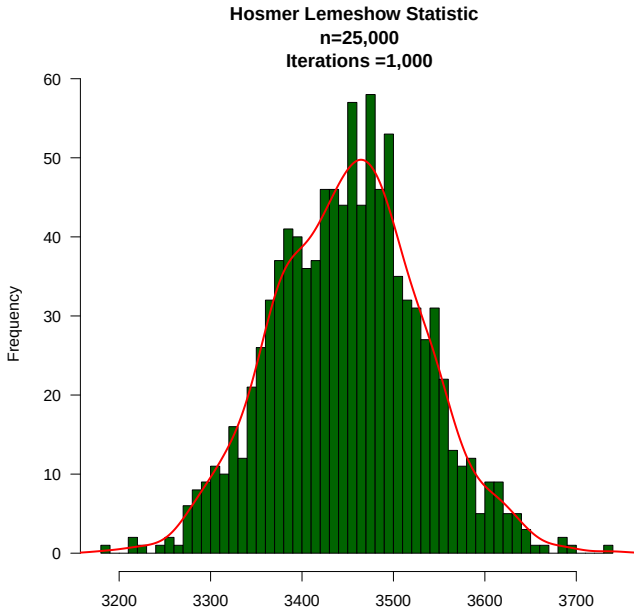

Figure 4: Distribution of the Hosmer-Lemeshow Test Statistic from simulation using n=25,000 with repetitions =1,000

6. Conclusion

There is much work that remains to find a viable solution for using the Hosmer-Lemeshow Test with large datasets. A greater understanding of the distribution of the the bootstrapped HLT need to be understood in order to develop, p-values and cutoffs that are theoretically justified. Future work needs to use simulated data where the true model is known and examine the performance of the proposed mHLT statistics. There needs to also be a push to get these tools into the hands of practitioners and applied statisticians. The default option for all statistical packages that the author is aware of, sets the number of groups equal to 10. This is only appropriate in small data situations. For the time being, practitioners should at least be adjusting the number of groups based on sample size where appropriate until a long term solution is found. The `determineG` function can be found in Logistic Regression Tools (lrt) Package[2], which is part of an R package currently under development. It will calculate the appropriate number of groups based upon how large that data are.

[2]https://github.com/Spoted21/lrt

References

[1] S. le Cessie and J. C. van Houwelingen. A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, 47(4):pp. 1267–1282, 1991.

[2] Venkat Chandrasekaran and Michael I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

[3] Dave Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *Math Challenges of the 21st Century*, 2000.

[4] D. W. HOSMER, T. HOSMER, S. LE CESSIE, and S. LEMESHOW. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980, 1997.

[5] Andrew A. Kramer and Jack E. Zimmerman. Assessing the calibration of mortality benchmarks in critical care: The hosmer-lemeshow test revisited *. *Critical Care Medicine*, 35(9):pp. 2052–2056, 2007.

[6] D. Y. LIN and D. ZENG. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):pp. 321–332, 2010.

[7] Prabasaj Paul, Michael L. Pennell, and Stanley Lemeshow. Standardizing the power of the hosmerlemeshow goodness of fit test in large data sets. *Statistics in Medicine*, 32(1):67–80, 2013.

[8] Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.

[9] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 4 2011.