Received: 14 October 2019 | Accepted: 5 November 2019

DOI: 10.1111/biom.13251

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

DISCUSSION

Discussion on "Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test" by Giovanni Nattino, Michael L. Pennell, and Stanley Lemeshow

Ivy Liu[1] | Daniel Fernández[2,3]

[1]School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

[2]Serra Húnter fellow, Department of Statistics and Operations Research, Polytechnic University of Catalonia-BarcelonaTech, Barcelona, 08028, Spain

[3]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Madrid, 28029, Spain

Correspondence:
Ivy Liu, School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand.
Email: Ivy.Liu@vuw.ac.nz

Funding information
Marsden Fund, Grant/Award Number: E2987-3648; Ministerio de Ciencia e Innovación, Grant/Award Number: RTI2018-100927-J-I00

It is our pleasure to comment on this paper. We would like to congratulate Giovanni Nattino, Michael L. Pennell, and Stanley Lemeshow for their great contribution to the goodness-of-fit method. One of the advantages of the Hosmer-Lemeshow approach (Hosmer and Lemeshow, 1980) is the simplicity of the method. It is the most popular way to evaluate the goodness-of-fit of logistic regression models. Their paper proposes a modified Hosmer-Lemeshow method that has the same advantage of simplicity, but for large data sets. We could foresee the popularity of the proposed method and its generalizations.

Our commentary focuses on a possible generalization of the method from binary responses to ordered response categories, which are also common in practice such as the Likert scale for survey data, the pain scale in medical studies, and the Braun-Blanquet coverage scale for ecological data.

There are a variety of approaches to the modeling of ordinal data that properly respect the ordinal nature of the data. Liu and Agresti (2005) and Agresti (2010) described various approaches using a proportional odds structure such as the proportional odds version of adjacent categories logits, cumulative logits (McCullagh, 1980), and continuation ratio logits (McCullagh and Nelder, 1989). That structure makes a strong assumption on common odds ratios and this may be inadequate for some data. Our discussion focuses on the ordered stereotype model (Anderson, 1984) that is more flexible than the models with the proportional odds structure as a result of adding additional score parameters. We briefly formulate it below.

Let Y_i be an ordinal response with q categories for observation i, where $i = 1, \dots, n$. The ordered stereotype model (Anderson, 1984) for the probability that Y_i takes the category k is characterized by the following log odds:

$$\log\left(\frac{P\left[Y_i = k \mid \boldsymbol{x}_i\right]}{P\left[Y_i = 1 \mid \boldsymbol{x}_i\right]}\right) = \alpha_k + \phi_k \boldsymbol{\beta}' \boldsymbol{x}_i,$$
$$i = 1, \dots, n, \qquad k = 2, \dots, q, \tag{1}$$

where the inclusion of the following monotone nondecreasing constraint $0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1$ ensures that the response Y_i is ordinal.

Regarding this model, the comparison of two proposed goodness-of-fit tests is evaluated. One is based on an extension of Hosmer-Lemeshow test for ordinal responses (Fernández and Liu, 2016), called the HL-type test (S), and the other one is the modified version for ordinal data proposed by this paper (Giovanni Nattino, Michael L. Pennell, and Stanley Lemeshow), called the modified HL test (M). Both tests are described below.

TABLE 1 Parameter configuration in the simulation study for assessing the power of the test when there is an omitted quadratic term in a continuous covariate (Scenario 1) and there is an omitted interaction term between a continuous covariate and a dichotomous variable (Scenario 2), when magnitude of the model's misspecification is small

q	Scenario	Covariates	$\{\alpha_k\}$	$\{\phi_k\}$	n	$\{\beta_j\}$
3	1	$\mathcal{N}(5,3)$	$(0,-0.8,-1.2)$	$(0,1/2,1)$	25 000	(1,0.0001)
					50 000	(1,0.0055)
					100 000	(1,0.01)
					500 000	(1,0.009)
					1 million	(1,0.008)
	2	$\mathcal{N}(5,3)$ & Bern(0.5)			25 000	(0.5,0.5,0.02)
					50 000	(0.5,0.5,0.04)
					100 000	(0.5,0.5,0.05)
					500 000	(0.5,0.5,0.045)
					1 million	(0.5,0.5,0.04)
4	1	$\mathcal{N}(5,3)$	$(0,0.1,-0.8,-1.2)$	$(0,1/3,2/3,1)$	25 000	(1,0.0001)
					50 000	(1,0.0055)
					100 000	(1,0.01)
					500 000	(1,0.009)
					1 million	(1,0.008)
	2	$\mathcal{N}(5,3)$ & Bern(0.5)			25 000	(0.5,0.5,0.02)
					50 000	(0.5,0.5,0.04)
					100 000	(0.5,0.5,0.05)
					500 000	(0.5,0.5,0.045)
					1 million	(0.5,0.5,0.04)
5	1	$\mathcal{N}(5,3)$	$(0,-0.1,-0.8,-1.2,-1.6)$	$(0,1/4,2/4,3/4,1)$	25 000	(1,0.0001)
					50 000	(1,0.0055)
					100 000	(1,0.01)
					500 000	(1,0.009)
					1 million	(1,0.008)
	2	$\mathcal{N}(5,3)$ & Bern(0.5)			25 000	(0.5,0.5,0.02)
					50 000	(0.5,0.5,0.04)
					100 000	(0.5,0.5,0.05)
					500 000	(0.5,0.5,0.045)
					1 million	(0.5,0.5,0.04)

1 | HL-TYPE TEST (S)

The S test statistic follows closely the test proposed in Fagerland and Hosmer (2013) for the proportional odds model. The main difference with our test is that we take advantage of score parameters $\{\phi_k\}$ from the ordered stereotype model to determine a new spacing of the ordinal response categories. The steps to construct the proposed test are as follows:

1. Let $\theta_{ik} = P[Y_i = k \mid \boldsymbol{x}_i]$ from (1). Calculate the estimated probabilities $\widehat{\theta}_{ik}$ for each observation $i = 1, \ldots, n$ and response category $k = 1, \ldots, q$.
2. Compute the weighted score for each observation:
$$s_i = \sum_{k=1}^{q} v_k \times \hat{\theta}_{ik}, \qquad i = 1, \ldots, n, \tag{2}$$
where $v_1 = 1$, $v_q = q$, and $v_k = 1 + (q-1) \times \hat{\phi}_k$. Note that the $\{v_k\}$ in the range of $[1, q]$ are the rescaled ordinal scores for the response categories, calculated from the score parameter estimates $\{\hat{\phi}_k\}$ in [0,1].
3. We sort the observations ascending by the weighted scores $\{s_i\}$.
4. We create a partition into G groups based on the weighted scores $\{s_i\}$, such that each group contains n/G observations. Sorting according to the weighted score follows closely the sorting used in Fagerland and Hosmer (2013).

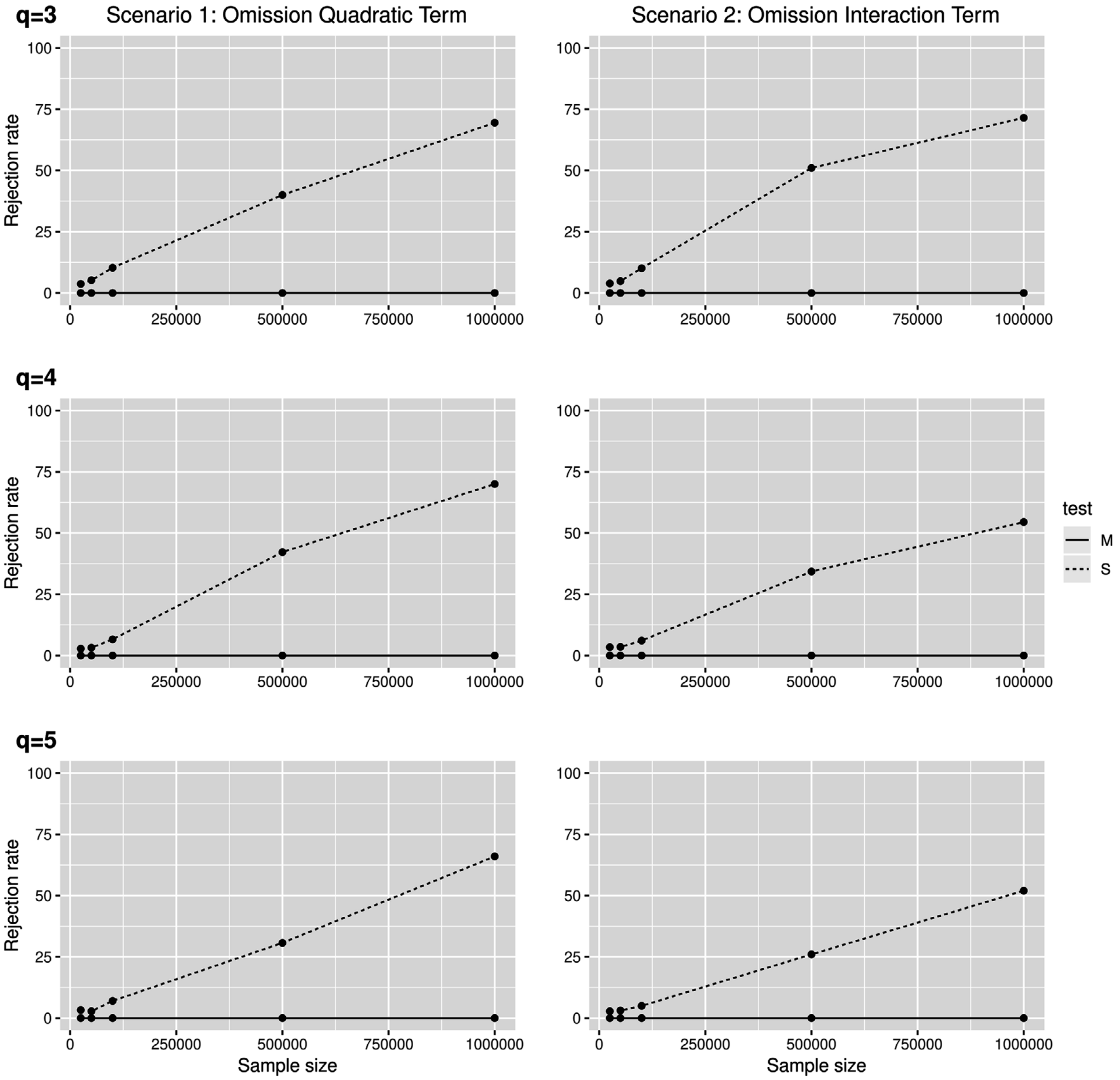

FIGURE 1 Graphical representation of the simulation scenarios for a level of significance of $\alpha = .05$, number of groups $G = 10$, and small magnitude of the model's misspecification. Percentage of rejections (out of 1000 simulated data sets) of the HL-type test (S, dashed lines) and the modified HL test (M, solid lines) over different sample sizes ($n = 25\,000$, 50 000, 100 000, 500 000, and 1 million) is shown. The first column of graphs corresponds to Scenario 1 (omitted quadratic term in a continuous covariate) and the second column of graphs corresponds to Scenario 2 (omitted interaction term between a continuous covariate and a dichotomous variable). Upper panels correspond to $q = 3$ ordinal responses, and middle and bottom panels to $q = 4$ and $q = 5$, respectively

5. Cross-classify the observations according to the G groups and the ordinal response categories to create a $G \times q$ contingency table. The observed frequencies $\{o_{gk}\}$ and the estimated expected frequencies $\{e_{gk}\}$ under the model are defined as:

$$o_{gk} = \sum_{v \in \Upsilon_g} I[y_v = k] \text{ and } e_{gk} = \sum_{v \in \Upsilon_g} \hat{\theta}_{vk},$$

$$\text{for } g = 1, \ldots, G, \quad k = 1, \ldots, q,$$

where Υ_g denotes the set of indices of the observations in group g and $I[A]$ is a binary indicator that takes value 1 if A is true and 0 otherwise.

6. Compute the Pearson χ^2 statistic S as:

$$\text{S} = \sum_{g=1}^{G} \sum_{k=1}^{q} \frac{(o_{gk} - e_{gk})^2}{e_{gk}}, \tag{3}$$

which follows a chi-squared distribution with $(G-2)(q-1) + (q-2)$ degrees of freedom under the null hypothesis of perfect fit.

TABLE 2 Power of the test (%) for the detection of an omitted quadratic term in a continuous covariate (Scenario 1) and for the detection of an omitted interaction term between a continuous covariate and a dichotomous variable (Scenario 2), when magnitude of the model's misspecification is small

				$n = 25\,000$			$n = 50\,000$			$n = 100\,000$			$n = 500\,000$			$n = 1$ million		
q	Scenario	Test	Covariates	α : 1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
3	1	S	$x_1 \sim \mathcal{N}(5,3)$	0.5	3.7	7.5	0.8	5.2	10.5	1.7	10.3	15.4	22.5	40.0	55.0	43.0	69.5	77.5
		M		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2	S	$x_1 \sim \mathcal{N}(5,3)$	0.7	3.9	8.6	0.7	4.8	9.2	2.7	10.1	18.7	25.7	51.0	63.7	50.0	71.5	82.5
		M	$x_2 \sim$ Bern(0.5)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	1	S	$x_1 \sim \mathcal{N}(5,3)$	0.6	2.8	5.8	1.0	3.2	6.2	1.6	6.6	11.9	20.0	42.2	54.7	46.0	70.0	81.0
		M		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2	S	$x_1 \sim \mathcal{N}(5,3)$	0.3	3.4	6.7	1.0	3.5	7.6	1.8	6.1	12.4	16.0	34.3	49.7	35.5	54.5	69.0
		M	$x_2 \sim$ Bern(0.5)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	1	S	$x_1 \sim \mathcal{N}(5,3)$	0.7	3.3	6.2	0.4	2.8	6.3	2.1	7.0	11.8	14.7	30.7	43.7	39.0	66.0	75.5
		M		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2	S	$x_1 \sim \mathcal{N}(5,3)$	0.7	2.8	5.4	0.4	3.1	6.5	1.9	5.0	10.4	10.0	26.0	36.7	28.0	52.0	61.5
		M	$x_2 \sim$ Bern(0.5)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: S is the HL-type test and M is the modified HL test with the fitted probabilities from the ordered stereotype model using $G = 10$ groups with small magnitude of the model's misspecification over different sample sizes ($n = 25\,000$, 50 000, 100 000, 500 000, and 1 million) and ordinal response categories $q = 3, 4, 5$.

2 | MODIFIED HL TEST (M)

The M test follows the same six steps described in the S test. If the model does not hold, the distribution of the test statistic (3) is noncentral chi-squared with $(G-2)(q-1)+(q-2)$ degrees of freedom and noncentrality parameter λ. Define the standardized noncentrality parameter of the test statistic (3) as

$$\epsilon = \sqrt{\frac{\lambda}{n}}.$$

The estimator of ϵ is

$$\hat{\epsilon} = \sqrt{\frac{\max\{\mathrm{S} - ((G-2)(q-1)+(q-2)), 0\}}{n}}.$$

Following the proposed approach by this paper in the case of large data:

1. Consider $\mathcal{H}_0$: $\epsilon \leq \epsilon_0$ versus $\mathcal{H}_1$: $\epsilon > \epsilon_0$ with the suggested ϵ_0 as

$$\epsilon_0 = \sqrt{\frac{\chi^2_{\lambda=0, df=(G-2)(q-1)+(q-2), \alpha=0.05} - ((G-2)(q-1)+(q-2))}{10^6}}.$$

2. Obtain the p-value by

$$1 - F_{\epsilon_0 n, ((G-2)(q-1)+(q-2))}(\mathrm{S}),$$

   where $F_{\epsilon_0 n, ((G-2)(q-1)+(q-2))}(\cdot)$ is the cumulative density function of a noncentral chi-squared distribution with noncentrality parameter $\epsilon_0 n$ and $(G-2)(q-1)+(q-2)$ degrees of freedom.

3 | SIMULATION STUDY

We set up a small simulation study in two of the scenarios described by this paper: omission of a quadratic term in a continuous covariate (Scenario 1) and omission of an interaction term between a continuous and a dichotomous covariate (Scenario 2), when magnitude of the model's misspecification is small. We simulated 1000 data sets from $\alpha_k + \phi_k(\beta_1 x_1 + \beta_2 x_1^2)$ (Scenario 1) and $\alpha_k + \phi_k(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$ (Scenario 2), where $x_1 \sim \mathcal{N}(5,3)$ and $x_2 \sim$ Bern(0.5). The values of the parameter setting were tuned such that all scenarios reached the same level of ϵ. Table 1 gives the detail of all parameters. We varied the sample size ($n = 25\,000$, 50 000, 100 000, 500 000, and 1 million) and the number of ordinal response categories ($q = 3, 4, 5$). For each data set, the ordered stereotype model with only main effects was fitted. We calculated the proportion of times that the null hypothesis was rejected at a 1%, 5%, and 10% level of significance for S (HL-type) and M (modified HL) tests.

4 | DISCUSSION

Figure 1 and Table 2 report the rejection rate of the two tests in all scenarios when magnitude of the model's misspecification is small. We observed a similar pattern of results for all scenarios. For the HL-type test, the rejection rate increases as the sample size increases for misspecified models. Moreover, the rejection rate slightly decreases as the number of ordinal response categories q increases because the degrees of freedom of the chi-squared distribution for the HL-type test statistic depend on q. Thus, this test has lower power for a larger q.

The modified HL test seems to be more conservative for the ordinal case than for the binary case because we obtained a zero rejection rate over all scenarios.

A more comprehensive simulation study must be set to make further conclusions. For instance, we could increase the level of ϵ and try different scenarios. Nevertheless, the method developed by this paper provided new insights to deal with big data for a wide range of models.

ACKNOWLEDGMENTS

This work was supported by the Marsden Fund (Award Number E2987-3648) administrated by the Royal Society of New Zealand and also supported by grant RTI2018-100927-J-I00 administrated by Ministerio de Ciencia e Innovación (MCI, Spain), by the Agencia Estatal de Investigación (AEI, Spain), and by the European Regional Development Fund FEDER (FEDER, UE). We are grateful to Dr Louise McMillan for helpful comments.

ORCID

Ivy Liu https://orcid.org/0000-0002-3152-2632
Daniel Fernández https://orcid.org/0000-0003-0012-2094

REFERENCES

Agresti, A. (2010) *Analysis of Ordinal Categorical Data*, 2nd edition. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.

Anderson, J.A. (1984) Regression and ordered categorical variables. *Journal of the Royal Statistical Society Series B*, 46, 1–30.

Fagerland, M.W. and Hosmer, D.W. (2013) A goodness-of-fit test for the proportional odds regression model. *Statistics in Medicine*, 32, 2235–2249.

Fernández, D. and Liu, I. (2016) A goodness-of-fit test for the ordered stereotype model. *Statistics in Medicine*, 35, 4660–4696.

Hosmer, D.W. and Lemeshow, S. (1980) Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9, 1043–1069.

Liu, I. and Agresti, A. (2005) The analysis of ordered categorical data: an overview and a survey of recent developments. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 14, 1–73.

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society*, 42, 109–142.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edition. London: Chapman & Hall.

How to cite this article: Liu I, Fernández D. Discussion on "Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test" by Giovanni Nattino, Michael L. Pennell, and Stanley Lemeshow. *Biometrics*. 2020;1–5. https://doi.org/10.1111/biom.13251