An Entropy-Based Measure for Assessing Fuzziness in Logistic Regression

Educational and Psychological
Measurement

2016, Vol. 76(6) 986–1004

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415623820

epm.sagepub.com

Brandi A. Weiss¹ and William Dardick¹

Abstract

This article introduces an entropy-based measure of data–model fit that can be used to assess the quality of logistic regression models. Entropy has previously been used in mixture-modeling to quantify how well individuals are classified into latent classes. The current study proposes the use of entropy for logistic regression models to quantify the quality of classification and separation of group membership. Entropy complements preexisting measures of data–model fit and provides unique information not contained in other measures. Hypothetical data scenarios, an applied example, and Monte Carlo simulation results are used to demonstrate the application of entropy in logistic regression. Entropy should be used in conjunction with other measures of data–model fit to assess how well logistic regression models classify cases into observed categories.

Keywords

data–model fit, logistic regression, classification

Introduction

In social sciences, regression is a data-analytic tool used to test hypotheses about relations between an outcome variable and a set of predictor variables. Multiple linear regression (MLR) is based on the assumption that the outcome variable is continuously measured on at least an interval-level measurement scale and (as the name implies) that variables are linearly related. On the contrary, many researchers wish to analyze models that predict categorical outcomes. Some possible examples of categorical outcomes used in social science research include whether or not a person

¹The George Washington University, Washington, DC, USA

Corresponding Author:

Brandi A. Weiss, The George Washington University, 2134 G. Street NW, Washington, DC 20052, USA.

Email: weissba@gwu.edu

will: survive or not survive, pass/fail a test, be admitted/rejected from a program, drop out/stay in school, and employment status. When an outcome variable is categorical in nature, assumptions such as interval-level outcomes and linearity are not met, thus MLR is an inappropriate analysis. However, in data scenarios such as these, Logistic regression can be used as a data-analytic technique to test hypotheses with a categorical outcome. Logistic regression is conducted using the logit transformation of the outcome variable. The logistic regression prediction equation for a person i in a model with k predictors is

$$\text{logit}(p_i)' = \ln\left(\frac{p_i}{1 - p_i}\right) = a + b_1X_1 + \dots + b_kX_k \quad (1)$$

where p_i is the probability of the outcome for person i defined as

$$p_i' = \frac{1}{1 + e^{-(a + b_1x_1 + \dots + b_kx_k)}} \quad (2)$$

a is the y -intercept, b_i is the unstandardized regression coefficient for an associated predictor, e is the constant 2.71828, and X_i is a categorical or continuous predictor variable.

Before model parameters can be interpreted, evidence must be provided that the model fits the data. In multiple regression this task is straightforward as the F-ratio can be used to test whether the model explains statistically significantly more variance than zero (i.e., $H_0: R^2 = 0$). In logistic regression, however, this task is nontrivial. There are many measures that examine whether or not a model reproduces the data and often model fit indices lead researchers to different conclusions.

Current measures of data–model fit in logistic regression are limited with regard to the lack of statistical significance tests, disagreement in conclusions between statistical significance tests, sensitivity to cut-points, sensitivity to the sample and sample size, and subjectivity of visual inspection of graphs, potential lack of statistical power, and potential ineffectiveness at describing the accuracy of classification. On its own, a single measure of data–model fit provide researchers with unique information about the model. Methodologists, however, recommend that applied researchers examine multiple measures to ensure a model fits the data.

The goal of the current study was to introduce an entropy-based measure that can be used to assess the extent to which a logistic regression model fits the data. This measure compensates for some of the limitations associated with current measures of data–model fit and is based on previous research in mixture modeling that has used a measure of entropy, which quantifies how well individuals are classified into latent classes. We suggest a small change to the entropy formula so that it can also be used to quantify the separation between predicted group membership when group membership is known. In this article, we explain the limitations of current measures of data–model fit in logistic regression, describe how entropy has been used as a measure of data–model fit in latent class analyses, introduce a variation of the entropy

formula that can be used for logistic regression models, and illustrate applied examples of the usefulness of entropy for logistic regression. Hypothetical data scenarios are used to exemplify how entropy captures unique information that is not contained in other measures in logistic regression to assess data–model fit. We then use entropy in an applied data example, present Monte Carlo simulation study results, and implications are discussed.

Measures of Model Fit in Logistic Regression

Logistic regression is commonly used as a measured-variable analysis to examine the relation between a predictor (or a set of predictors) and a categorical outcome variable. To determine whether a proposed model fits the data researcher's commonly rely on measures such as the likelihood ratio test, Hosmer–Lemeshow (HL) chi-square (χ^2) goodness-of-fit test, pseudo R^2 values, the percentage of cases correctly classified, receiver operating characteristic (ROC) curves, and histograms of predicted probabilities. The first two measures provide statistical significance tests of whether the model fits the data, while the later measures can be thought of as effect sizes and measures of data–model fit. ROC curves and histograms of predicted probabilities and outcomes are methods of visual inspection. Each of these measures is briefly described below and limitations are discussed.

Deviance ($-2LL$)

In logistic regression analysis, deviance ($-2LL$) is a mathematical quantity that represents the lack of fit of a model by comparing a fitted model to a saturated model (Cohen, Cohen, West, & Aiken, 2003). As such, it represents how much observed values differ from expected values. The likelihood ratio is a ratio of the maximum likelihood of one model compared with another model; deviance is calculated by multiplying the log of the likelihood ratio by (-2) . Deviance values of zero represent perfect fit, while large values are indicative of poor fit. Deviance does not necessarily follow a chi-square distribution, thus it is more useful for model comparison than to interpret on its own (O'Connell & Amico, 2010; Peng, So, Stage, & St. John, 2002).

The Likelihood Ratio χ^2 Test

The likelihood ratio chi-square test (LR χ^2) uses deviance to compare a smaller model (often a baseline or intercept-only model) to a larger model. The change in log likelihood is computed by taking the difference in the $-2LL$ values between the two models. The difference in $-2LL$ follows a chi-square distribution and thus the value can be compared to a chi-square critical value. A statistically significant result indicates that the model with more predictors fits the data statistically significantly better than the model with fewer predictors. Benefits of this test include its usefulness in determining whether a theoretical model fits better than a baseline model,

and it can be used as a statistical test to compare nested models. The LR χ^2 is a limited measure of data–model fit when data are sparse (often a result when continuously measured predictors are used) or when the sample size is large.

Hosmer–Lemeshow χ^2 test

The HL χ^2 test is also a type of goodness-of-fit test useful when data are sparse. This statistic divides predicted probabilities into 10 groups (deciles) based on percentile ranks and a Pearson χ^2 is calculated to compare the observed and expected frequencies. Nonstatistically significant results are indicative of good data–model fit. This statistic is considered a conservative test (i.e., low Type I error rates; Peng et al., 2002) and is useful when continuous predictors are used. The measure is criticized for its lack of power and thus only recommended when $n > 400$ (Hosmer & Lemeshow, 2000). In contrast, chi-square tests are generally very sensitive to sample size and thus with large n 's they frequently produce statistically significant results (indicating poor data–model fit). Thus, there is no consensus of necessary sample size for ideal power for this test. Additionally, the measure is sensitive to cut-points used to create deciles and thus the numeric HL χ^2 value can vary across software programs (Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997; Peng et al., 2002).

Pseudo R^2 (Variance Explained) Measures

Several pseudo variance explained measures are sometimes used in logistic regression as effect size measures for how well a model fits the data (e.g., McFadden's R^2 , Cox & Snell, 1989; Nagelkerke, 1991). These are meant to be interpreted similarly to multiple R^2 in ordinary least squares (OLS) regression, but are considered pseudo R^2 measures because the variance in categorical outcomes differs from the variance in continuous outcomes (Lomax & Hahs-Vaughn, 2012). Although these measures are meant to range from 0 to 1 (with 1 being indicative of better data–model fit) it is mathematically impossible for the Cox and Snell (1989) measure to equal 1.0, and recommended cut-off values do not currently exist for any pseudo-variance explained measure. Of these, McFadden's R^2 index makes the most intuitive sense because its interpretation is most similar to R^2 in multiple regression, however, it is not available in many commercial statistics software programs, and thus must be computed by hand making it an undesirable measure for some applied researchers (O'Connell & Amico, 2010; Peng et al., 2002). Additionally, there is disagreement among researchers in regards to which of these pseudo R^2 values is preferable to report (O'Connell & Amico, 2010).

The Percent of Cases Correctly Classified

Many researchers rely on classification tables that contain cross-tabulations of the observed group membership with predicted group membership. In binary logistic

regression, cases with predicted probabilities of .5 or greater are typically classified into one group, whereas predicted probabilities less than .5 result in predicted group membership in the second group (although if desired, researchers could select a cut-point other than .5). Based on the cross-tabulations table, the percent of cases correctly classified can be computed and used as a measure of data-model fit. In the binary logistic regression case, classification rate values range from 50% to 100% correct classification, where larger values are indicative of better data-model fit.

Percent of correct classification is often criticized as being an ineffective measure of data-model fit due to the large misclassification rate for groups with smaller observed frequencies (particularly when the sample size for a given group is small,) potential poor classification rates for one or more groups and thus underrepresentation of these groups within the classification rate, and not incorporating base rates or chance classification (Lomax & Hahs-Vaughn, 2012; O'Connell & Amico, 2010). Furthermore, situations may arise where a model is considered to fit the data well, but the classification rate may be low.

Receiver Operating Characteristic Curves

There are also several methods of visual inspection that can be used to determine whether or not a model fits the data. ROC curves compare sensitivity (the percent of true positives, i.e., cases that were correctly classified as an event) to 1-specificity (the percent of false positives, i.e., cases that were misclassified as an event). Each sensitivity-and-(1 - specificity) pair is plotted across probability cut-points ranging from 0 to 1. Thus, the ROC curve indicates the percent of cases that need to be misclassified in order to correctly classify a given percent of cases. ROC curves can also be useful in binary logistic regression for identifying cut-points other than .50.

Area Under the Receiver Operating Characteristic Curve

The major limitation of the ROC curve is that it only provides visual inspection and thus its interpretation for data-model fit is subjective. Researchers can, however, use ROC curves to calculate the area under the curve (AUC). The AUC quantifies the curve by reporting a single numeric value that can be used as a measure of data-model fit. Conceptually, AUC is the probability of concordance, which is the probability that a randomly selected positive case will score higher than another randomly selected negative case. The closer the curve is in the upper-left corner, the larger the AUC. For binary logistic regression, AUC ranges from .5 to 1.0 where higher values are indicative of better data-model fit. AUC produces identical results to the Wilcoxon tests of ranks (Streiner & Cairney, 2007). As a rule of thumb, AUC values between .5 to .7 are considered low; .7 to .9 are moderate, and values greater than .9 are high (Streiner & Cairney, 2007). A *t* test can be computed to determine if the AUC is statistically significantly different from 0.5. Statistically significant AUC

values indicate that the model fits better than chance. Some researchers report a 95% confidence interval around the AUC estimate.

There are many problems using ROC curves and AUC as measures of data–model fit. First, the sensitivity and specificity values depend on where the cut-points occur, thus AUC can be calculated an infinite number of ways. Additionally, some researchers caution against using ROC curves for model comparison because they do not take variance into account and ROC curves may cross thus they are not directly comparable (Fawcett, 2005; Streiner & Cairney, 2007).

Histogram of Predicted Probabilities

Finally, the histogram of predicted probabilities (also referred to as a classification plot) and outcomes can be used to examine the quality of prediction. This histogram provides the percent of false positives, percent of false negative, and percent of correct classification in a visually informative graph. To indicate well-differentiated prediction, classification plots should have the following characteristics: (1) U-shape (rather than normally distributed), (2) cases clustered at each end (i.e., further away from the cut-point, usually .5 in binary logistic regression), and (3) minimal errors in classification (i.e., few false positive and few false negative). The histogram complements the classification rates measure by graphically depicting the percent of correct classifications as well as the separation between groups. Unfortunately, there is not a single numeric value that quantifies the information contained within this histogram in the way that AUC quantifies the ROC curve.

Limitations of Current Measures

Methodologists recommend that applied researchers use a combination of data–model fit measures to evaluate logistic regression models (Hosmer & Lemeshow, 2000; O’Connell & Amico, 2010). Together, the measures described may be useful for evaluating whether or not a model fits the data, however, all these measures have limitations and may not provide a complete picture of whether or not the model fits the data and where potential misfit may be occurring. Additionally, while the histogram of predicted probabilities plot contains a lot of information and is helpful in determining the quality of prediction, the usefulness of the plot is limited in that it only provides researchers with visual inspection, which is subjective and susceptible to experimenter bias. Thus a quantitative index that summarized the plot in a single numeric value is desirable because it would reduce experimenter bias and provide standardization across studies.

Entropy as a Measure of Model Fit in Latent Class Analysis

One measure of data–model fit that has been used in latent class analysis (LCA) is entropy. Entropy is a classification-based approach that has been used to help

researchers determine the number of latent classes that exist in a dataset (Celeux & Soromenho, 1996; Clark & Muthén, 2009; Henson, Reise, & Kim, 2007; Pastor & Gagne, 2013). Entropy was first introduced by Ramaswamy, Desarbo, Reibstein, and Robinson (1993) as a measure in LCA to assess “fuzziness” (i.e., the lack of separation) between classes. In LCA the entropy index is calculated as

$$\text{Entropy} = 1 - \frac{E_k}{N \ln K}, \quad (3)$$

where N is the total sample size, K is the number of latent classes, and E_k is defined as

$$E_k = \sum_{i=1}^N \sum_{k=1}^K (-p_{ik}^* \ln p_{ik}), \quad (4)$$

where $0 \leq E_k \leq 1$ and p_{ik} represents the conditional posterior probabilities calculated for each observation, i , and represents the probability of membership in each of the K classes. As posterior probabilities move further from a cut-point used to differentiate between classes (in a two-class model this cut-point is often set at .50), the “fuzziness” of class membership decreases due to the increased separation of classes, and thus the value of entropy also increases to reflect this better separation. Thus, higher values of entropy are more desirable because they are indicative of good classification quality (i.e., good separation and thus lack of fuzziness between classes). More specifically, perfect entropy values of 1 indicate all persons are perfectly classified (i.e., all posterior probabilities are extreme values of 0 and 1), whereas values of 0 represent complete fuzziness between the classes (i.e., the posterior probabilities are at the cut-point used to differentiate class membership).

Applying Entropy in Logistic Regression

Logistic regression is similar to LCA in that the outcome variable (i.e., class/group membership) is categorical. However, logistic regression differs from LCA because in the former group membership is known. Thus, Equations (3) and (4) can be adapted to evaluate the amount of “fuzziness” between predicted group membership in logistic regression models. To do so, a modification needs to be made in how p_{ik} is defined in Equation (4). Specifically, since class membership is known in logistic regression, p_{ik} represents the predicted probabilities calculated for each observation in each of K groups rather than the conditional posterior probabilities. For clarity E_k^* is used to define entropy for cases of predicted probabilities and is defined as

$$E_k^* = \sum_{i=1}^N \sum_{k=1}^K (-\hat{p}_{ik}^* \ln \hat{p}_{ik}) \quad (5)$$

Table 1. Classification Rates Verses Entropy Across Four Hypothetical Scenarios.

Case No.	Observed group	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
		<i>P'</i>	<i>K'</i>	<i>P'</i>	<i>K'</i>	<i>P'</i>	<i>K'</i>	<i>P'</i>	<i>K'</i>
1	1	.501	1	.999	1	.501	1	.999	1
2	1	.501	1	.999	1	.499	0	.001	0
3	0	.499	0	.001	0	.501	1	.999	1
4	0	.499	0	.001	0	.499	0	.001	0
Classification rate		100%		100%		50%		50%	
Entropy		0.00		1.00		0.00		1.00	

Note. *P'* = predicted probability; *K'* = predicted group membership.

where $0 \leq E_k^* \leq 1$ and $\hat{p}_{ik}$ represents the predicted probability of membership for each observation, *i*, in each of the *K* groups. Similarly to entropy in LCA, E_k^* values close to 1 are indicative of good classification quality.

Hypothetical Scenarios

Table 1 contains four hypothetical scenarios that exemplify how entropy provides a unique measure of data–model fit for binary logistic regression analyses. Here, data are shown for the most simplistic four-case scenario and observed group membership is provided for these four cases. Correspondingly, Figure 1 graphically displays the predicted probabilities of these four cases in a histogram. Recall that in binary logistic

regression .5 is typically used as a cut-point for predicted group membership. Ideally, predicted probabilities for each case in each group would approach 0 or 1 and thus be indicative of minimal “fuzziness” and better data–model fit.

In the first scenario, the predicted probabilities are very close to the .5 cut-point value. This scenario depicts perfect classification (classification rate of 100%), however, entropy has a value of 0. Although we have perfect classification in this scenario, the predicted probabilities for all cases are very close to the .50 cut point, which means the quality of classification is low and thus group membership predictions may not generalize back to the population. This instability of classification is not represented in the classification rate. The entropy index, however, is very low and thus informs us that the quality of classification is poor, thus entropy provides us with unique information not contained within the classification rate. Furthermore, entropy is advantageous in that it quantifies the information contained in the histogram of predicted probabilities (shown for Scenario 1 in Figure 1) in a single numeric value.

In the second scenario, predicted probability values approach 1 and 0, resulting in perfect classification. In this scenario, entropy has a high value of 1. The combination of these fit indices suggests that for this scenario the theoretical model fits the data well (high classification rate) and the quality of those classifications is strong (high entropy value). Scenario two is the most desirable scenario.

In terms of data–model fit, Scenario 3 is undesirable because the classification rate is only 50% (the lowest possible classification rate that can be observed in the binary case) and entropy is a low value of 0. In Scenario 3, both the classification rate and entropy indicate that the model does not fit the data well. In the final scenario the classification rate is poor (50%), however, entropy is perfect (1.0). In the fourth scenario, we have the poorest possible classification and are nearly certain in that classification as indicated by perfect entropy.

When comparing Scenario 1 with Scenario 2 we come to the same conclusions based on overall classification but in the second scenario we are extremely confident in our conclusions. Furthermore separation, as indicated by entropy, adds strength to any claim made regarding fit of the theoretical model. That is we might expect this to be a measure of how likely our results would replicate in the future. Low entropy does not seem to provide confidence that our categories will be separated the same way should we draw another sample from the population. Given these two scenarios a researcher may desire to maximize entropy. However, when we compare Scenario 3 with Scenario 4 we are also extremely confident in our conclusions regarding the later scenario, Scenario 4, however we now have poor classification. In this case the researcher would likely want to minimize entropy when coupled with poor classification.

Empirical Application

In this section, a dataset for which logistic regression analysis is appropriate to predict a dichotomous outcome is described. SPSS version 22.0 was used to conduct

binary logistic regression analyses using maximum-likelihood estimation. Entropy was calculated and checked using SPSS, Microsoft Excel, and R. Model-fit statistics for three competing models are presented and discussed.

Data and Logistic Regression Models

To illustrate the application of entropy we used the *Titanic* dataset. The *Titanic* was a British passenger ship that collided with an iceberg and sank in the North Atlantic Ocean on April 15, 1912. This dataset has previously been used to illustrate statistical analyses of categorical data (e.g., Dawson, 1995; Harrell, 2001). The dataset is based on the passenger information originally published by Eaton and Haas (1994) and has been collectively edited by a variety of researchers in the Internet community. The total number of passengers onboard differs across sources but most agree there were approximately 1,300 passengers on board. The dataset used in the current study contains information from 1,309 passengers and does not include crew members. Due to missing data on independent variables, 1,045 passengers were used for Models 1 and 2, and 1,308 passengers were used for Model 3. Missing data were not imputed due to the illustrative application purpose of this article.

The dichotomous outcome variable was survival (did not survive = 0, survived = 1). Three models were compared using the following six independent variables: fare (measured in British pounds), age (in years), sibsp (the number of siblings and spouses aboard), parch (the number of parents and children aboard), gender (males = 0, females = 1), and whether or not passengers were on a lifeboat (0 = no lifeboat, 1 = lifeboat). Model 1 had four predictors: fare, age, sibsp, and parch. Model 2 included all predictors in model one, plus gender. Model 3 had the single predictor of lifeboat. Because Model 2 was nested in Model 1, Model 2 was hypothesized to have better data-model fit than Model 1. As survival was highly dependent on whether or not passengers had a lifeboat (95.4% of passengers with a lifeboat survived), Model 3 was hypothesized to be the best-fitting model.

Results

Table 2 contains the data-model fit indices for the three models. The LR χ^2 tests indicate that all models fit statistically significantly better than an intercept-only model ($p < .001$). On the contrary, the HL χ^2 tests for Models 1 and 2 were statistically significant indicating that these models did not fit well (i.e., observed frequencies were statistically significantly different than expected frequencies). An HL χ^2 test could not be computed for Model 3 given the categorical nature of the single predictor, and thus lack of values on the predictor variable available to create deciles. Figure 2 contains ROC curves for all models. Visual inspection of the ROC curves indicated that each model visually appeared to be better than chance. The AUC for each ROC curve was statistically significant indicating that the models classified groups statistically significantly better than chance. While visual inspection of the

Table 2. Data–Model Fit Indices for Three Logistic Regression Models Using the Titanic Dataset.

	AUC									
	95% CI					Cox and Snell R^2	Nagelkerke R^2	Classification rate (%)	% Δ in classification rate ^a	Entropy
	LR $\chi^2 p$	HL $\chi^2 p$	AUC	LB	UB					
Cut-off criteria	<.05	$\geq .05$								
Model 1	<.001*	<.001*	.696	.663	.729	.092	.124	67.0	7.9	.09
Model 2	<.001*	.034*	.825	.799	.851	.298	.402	77.5	18.4	.28
Model 3	<.001*	NA	.971	.960	.983	.668	.908	97.6	35.8	.84

Note. LR = likelihood ratio; HL = Hosmer and Lemeshow; AUC = area under the receiver operating characteristic curve; CI = confidence interval; LB = lower bound; UB = upper bound; NA = not applicable.

^aIn comparison to the blind model. Blind model classification rates were 59.1% for Models 1 and 2, and 61.8% for Model 3.

* $p < .05$.

Figure 2. Receiver operating characteristic (ROC) curves for three models using the Titanic survival dataset.

ROC curves indicated that fit improved with each model, caution should be made when using ROC curves and AUC to make model comparisons (Fawcett, 2005; Streiner & Cairney, 2007). Effect sizes including pseudo R^2 values, classification rates, and AUC are also shown in Table 2 for each model. As hypothesized, effect sizes improved with each model, indicating that Model 2 fit better than Model 1, and Model 3 fit better than both Models 1 and 2. As hypothesized, entropy also improved with each model.

If a researcher were only to test Models 1 and 2, he or she may conclude that based on currently existing measures of data–model fit and effect sizes Model 2 fit the data best, and thus continue to interpret Model 2 as being the “correct” model. However, entropy provides us with more information that we do not currently get from existing measures of model fit. On further investigation, the entropy value for Model 2 was quite low at 0.28. This is an indication that while Model 2 is better than a blind model for classifying cases into groups, there is little separation of predicted probabilities for group membership. Thus, while Model 2 fits the data, there is some fuzziness of class membership. Figure 3 contains the histograms of predicted probabilities for the three models tested. As shown in Figure 3, the classification plot for Model 2 is U-shaped; however, the two classes do not have much separation, particularly for those predicted to survive (on the right side of the classification plot), and there are a

number of false positives and false negatives. Thus, entropy has captured the “fuzziness” of this classification plot whereas other measures of data–model fit did not. Of the three models, Model 3 had the most desirable classification plot in which the graph is U-shaped, there is complete separation, and minimal classification errors.

Monte Carlo Simulation Study

The purpose of this study was to examine how entropy varies in simple binary logistic regression models. We used Monte Carlo simulation to generate sample data from known population parameters to evaluate the performance of entropy in comparison to other measures of data–model fit.

Data Generation

All variables were simulated to come from a population in which

$$\text{logit}(p_i)' = 0 + b_1x_1 \quad (6)$$

where p_i was the probability of the outcome for person i defined as

$$p_i' = \frac{1}{1 + e^{-(0 + b_1x_1)}} \quad (7)$$

where a was the y -intercept set at zero, b_i was the unstandardized regression coefficient for a single continuously measured predictor, e was the constant 2.71828, and x_i was the single continuously measured predictor.

We generated data to represent three scenarios meant to be similar to hypothetical Scenarios 1, 2, and 3 (respectively), presented earlier in this article. Scenario 1 was generated to have high classification and low entropy (separation of groups); Scenario 2 was the most ideal with high classification and high entropy; and Scenario 3 was the least ideal with low classification and low entropy. We varied the group means and standard deviations in order to reflect these three scenarios. Table 3 contains a summary of the Monte Carlo conditions and population specifications.

Scenario 2 was generated using *simple structure* as defined by Nylund, Asparouhov, and Muthén (2007), in which large mean differences between groups are simulated so that the model will discriminate well between the groups. Scenario 2 was specified to have high classification and high separation of groups. Specifically the observed means of each group were set at -2 or $+2$ with a standard deviation of 1. Scenario 3 was generated using *complex structure* in which the model is not able to distinguish well between group membership. Thus, for Scenario 3, group means were set at -1 or $+1$ with a standard deviation of 4. Finally, Scenario 1 was generated with the goal of showing the potential utility of entropy in comparison with other fit indices. Specifically, data for Scenario 1 were generated to have high classification but low separation between group membership (i.e., low entropy). For Scenario 1, group means were set at -1 or $+1$ and a standard deviation of 1.

Three different sample sizes were selected, which corresponded to $n = 100, 500$, and 1000 . The sample sizes were chosen based on past simulation studies in logistic regression and latent class analysis. Equal class sizes were used as is typical for designs with simple structure (Nylund et al., 2007). For simplicity, we kept equal

Table 3. Summary of Monte Carlo and Population Specifications.

Population specification		
Sample size	100; 500; 1,000	
Replications	10,000	
Model parameters	Class 1	Class 2
Class size	50%	50%
Scenario 1 (high classification, low entropy)		
Mean	−1	1
SD	1	1
Scenario 2 (high classification, high entropy)		
Mean	−2	2
SD	1	1
Scenario 3 (low classification, low entropy)		
Mean	−1	1
SD	4	4

class sizes for all conditions, which yields a 50% classification rate for baseline models.

In summary, the simulation design was a fully crossed 3 (sample size) $\times$ 3 (scenarios) factorial design resulting in 9 data generation conditions. All data were simulated to come from a normal distribution. Data were simulated and analyzed in SAS 9.3. Each condition was replicated 10,000 times. Binary logistic regression using maximum-likelihood estimation was used to obtain predicted probabilities. Entropy and other commonly used measures of data–model fit were calculated and compared.

Simulation Results

Results from the simulation study are shown in Table 4. Average fit statistics across the 10,000 replications are presented as well as the standard deviations of estimates for classification rates and entropy. As is evident from the table, sample size did not have an impact on results across scenarios, except that when the sample size was 100, the standard deviation around entropy values and classification rates was larger than other conditions, indicating less stability in these values.

For all scenarios, the average LR χ^2 value was statistically significant, the HL χ^2 values were not statistically significant, and the AUC values were above average and statistically significant, indicating on average these models fit the data well across three scenarios and conditions. Similarly, on average the classification rates were higher than that for the blind model (blind model was 50% due to equal class sizes) indicating that on average the model classified participants into groups better than no model at all. The pseudo R^2 values indicated that on average Scenarios 1 and 2 had good data–model fit, but the models did not fit the data well for Scenario 3.

Table 4. Data—Model Fit Indices From Simulation Study.

Population-generating simulated condition				Average observed parameter estimates								
Classification rate	Entropy	N	LR $\chi^2 p$	HL $\chi^2 p$	AUC		Cox and Snell R^2	Nagelkerke R^2	Classification rate		Entropy	
					AUC	p			Mean	(SD)	Mean	(SD)
Cutoff criteria												
Scenario 1												
			<.05	≥ .05	<.05							
High	Low	100	<.001*	.513	.921	.019*	.492	.656	84%	(.03)	.491 (.06)	
High	Low	500	<.001*	.504	.921	.038*	.491	.654	84%	(.01)	.487 (.03)	
High	Low	1000	<.001*	.499	.921	.002*	.490	.654	84%	(.01)	.487 (.02)	
Scenario 2												
High	High	100	<.001*	.932	.998	.008*	.721	.961	98%	(.01)	.921 (.04)	
High	High	500	<.001*	.864	.998	.017*	.718	.958	98%	(.01)	.914 (.02)	
High	High	1000	<.001*	.819	.998	.001*	.718	.958	98%	(.01)	.914 (.01)	
Scenario 3												
Low	Low	100	<.001*	.488	.638	.006*	.063	.084	59%	(.04)	.047 (.03)	
Low	Low	500	<.001*	.497	.638	.012*	.060	.080	60%	(.02)	.045 (.01)	
Low	Low	1000	<.001*	.495	.638	.001*	.059	.079	60%	(.01)	.044 (.01)	

Note. LR = likelihood ratio; HL = Hosmer and Lemeshow; AUC = area under the receiver operating characteristic curve. Classification rates should be compared with 50% blind model classification rate due to equal class sizes across conditions.

* $p < .05$.

The entropy values shown in Table 4 provide us with unique information. For Scenario 2, entropy leads us to a similar conclusion as the other measures of data–model fit; that is, on average the models fit the data well. For Scenario 3, entropy provides similar information as the pseudo R^2 measures, on average the models do not fit the data well. Scenario 1 is where entropy provides us with useful information not captured by any other measure of data–model fit. For Scenario 1, the value of entropy is in the middle (around .49). For latent class analysis, entropy values around .80 are often desirable to indicate good data–model fit. For Scenario 1, the average entropy values for these models fall well below this value, indicating that although the models fit better than a blind model and have good classification rates, there is little separation between the predicted probabilities values. Entropy tells us that Scenario 1 results in poorer than desired quality of classification, and perhaps the results may not replicate well to other samples.

Discussion

Methodologists recommend that applied researchers use a combination of data–model fit measures. In logistic regression, currently existing data–model fit indices are criticized for being sensitive to sample size, subjectively based on visual inspection of graphs, potentially lacking statistical power, being ineffective at describing the accuracy of classification, and when used in conjunction frequently lead to conflicting conclusions about whether or not the model fits the data. In the current study, we described a revised measure of entropy and showed how it can be useful in assessing the quality of logistic-regression models while complementing data–model fit indices already used. Using hypothetical scenarios, empirical data from the *Titanic*, and simulation study results we showed how entropy captures unique information that is not reflected in other measures of data–model fit currently used in logistic regression.

In the empirical data example, a researcher who relied on currently existing measures of model–fit statistics and classification rates, may conclude that both Models 2 and 3 fit the *Titanic* data. However, entropy provided us with a unique perspective into these data, and indicated that Model 2 had poor group separation. Entropy mathematically penalizes researchers for having borderline predicted probabilities. Thus, low values of entropy indicate some amount of misclassification or “fuzziness” of prediction. Similarly, in our simulation study conditions in which there was high classification but low separation of classes, most measures of data–model fit indicated that the model fit the data well. Entropy, however, was the only measure of data–model fit that indicated these models may be less than ideal.

This is a generalizability and replicability issue. Researchers are often reluctant to rely on classification rates because a model with favorable classification rates but poor separation (as indicated from visual inspection of the classification plot) may not replicate well in other studies. Entropy, however, captures information about borderline cases that classification rates do not. Specifically, borderline false positives

and false negatives may not generalize back to the population due to the instability of predicted probabilities. A slight shift in the predicted probabilities would result in differences in predicted group membership for these borderline cases. A model with good separation, however, would yield more stable results that generalize well to the population and replication studies. Thus, entropy may be a measure of the strength of the model to replicate categories. It is important to note that replicability does not necessarily mean that predicted classification is correct.

Similarly, entropy as a quality measure may be understood as a measure of certainty. That is, the greater the value of entropy, the more certain we are regarding the decision of classification for each case. When this interacts with classification we sometimes find we are certain regarding classifications that are incorrect. In LCA, categories are unknown, thus overreliance on entropy could result in misleading confidence about class membership. However, in logistic regression group membership is known, and thus the degree of classification is observable, thus entropy may be a stronger quality data–model fit index in logistic regression than it is in LCA.

Separation occurs with near perfect prediction (O’Connell & Amico, 2010). At first thought, separation seems like a desirable condition. O’Connell & Amico (2010), however, caution that while separation may seem desirable, it may prohibit researchers from understanding the effects of individual predictors within a model. Similar to conditions with sparse data and multicollinearity, separation results in large standard error estimates and regression coefficient estimates. Since entropy is a measure of the separation of groups/classes, there may be a point in which large values of entropy are undesirable because they result in biased estimates of standard errors and regression coefficients. Future studies should investigate cut-points for desirable values of entropy.

While the current study focused on the binary logistic regression model, entropy is versatile enough that it can be extended to other analyses with categorical outcomes (such as polytomous and ordinal logistic regression), and it can be used for model comparison (both nested and non-nested). In conclusion, entropy provides unique information that may not be captured by other measures of data–model fit currently used in logistic regression. Models of fit only provide partial information on their own (Long, 1997). While entropy may provide unique information it should be used in conjunction with other measures of data–model fit in logistic regression.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13, 195-212.
- Clark, S. & Muthén, B. (2009). *Relating latent class analysis results to variables not included in the analysis*. Retrieved from <http://www.statmodel.com/download/relatinglca.pdf>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London, England: Chapman & Hall.
- Dawson, R. M. (1995). The "unusual episode" data revised. *Journal of Statistics Education*, 3(3). Retrieved from <http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html>
- Eaton, J. P., & Haas, C. A. (1994). *Titanic: Triumph and tragedy*. New York, NY: W. W. Norton.
- Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.
- Harrell, F. E. (2001). *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer-Verlag.
- Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: a comparison of relative model fit statistics. *Structural Equation Modeling*, 14, 202-226.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.
- Hosmer, D. W., 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.
- Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *An introduction to statistical concepts*. 3rd ed. New York, NY: Routledge.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691-692.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.
- O'Connell, A. A., & Amico, K. R. (2010). Logistic regression. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 221-239). New York, NY: Routledge.
- Pastor, D. A., & Gagne, P. (2013). Mean and covariance structure mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 197-224). Charlotte, NC: Information Age.
- Peng, C. Y., So, T. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education*, 43, 259-293.
- Ramaswamy, V., Desarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science*, 12, 103-124.
- Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristic curves. *Canadian Journal of Psychiatry*, 52, 121-128.