# An Overview of the Assessment of Logistic Regression Models

Justin Shang* Timothy J. Robinson† Shaun S. Wulff‡

**Abstract**

The logistic regression model is frequently used in many practical applications to fit a binary response. Model specification depends upon a number of issues including response selection, link specification, and the choice of predictors. Model evaluation includes model selection, predictive ability, and goodness-of-fit. As a result, the art of logistic regression modeling involves many choices and multiple criteria for the data modeler to consider. Particular emphasis will be given to a thorough review of the model selection procedures and the goodness-of-fit testing. In logistic regression, goodness-of-fit assessments sometimes can be challenging, depending on the covariates in the model and the number of covariate patterns. Goodness-of-fit tests can involve chi-square based tests, raw residuals, and transformed residuals. We detail these approaches for assessing the quality of logistic regression models.

**Key Words:** goodness-of-fit, model selection, chi-square, residual, binary response

## 1. Introduction to logistic regression

Logistic regression has become a useful tool since the 1950's. Later, in the 1970's, logistic regression and other models (e.g., Poisson regression) derived from the exponential family of distributions were formalized into a generalized linear model (GLM) framework (Nelder and Wedderburn (1972)). McCullagh and Nelder (1989) and Agresti (1990) further developed the GLM framework. The concepts of logistic regression will be briefly introduced as well as challenges in model selection and assessing the model-fit.

### 1.1 Logistic regression

Consider a $N \times 1$ vector of binary responses $\underline{Y} = (Y_1, Y_2, \ldots, Y_N)'$, where $Y_i$ is coded as 1 or 0 for $i = 1, 2, \ldots N$. For the convenience purpose, here we refer coding of 1 as positive, and 0 as negative, although 1 sometimes may index success or presence, and 0 may index failure or absence. Consider an observed $p \times 1$ vector of regressors for observation $i$ denoted $\underline{x}_i = (1, x_{1i}, ...x_{ki})'$ with $p = k + 1$. Consider a corresponding $p \times 1$ vector of regression coefficients $\underline{\beta} = (\beta_0, \beta_1, ...\beta_k)'$, whereby the probability of a positive outcome $(\pi_i)$ is assumed to depend upon the vector of regressors $(\underline{x}_i)$ and the regression coefficients $(\underline{\beta})$, i.e., $\pi_i = \pi(\underline{x}_i, \underline{\beta})$. Let the response $(Y_i)$ has a *Bernoulli* distribution with mean and variance given by

$$E(Y_i) = \mu_i = \pi(\underline{x}_i, \underline{\beta}), \tag{1}$$

$$var(Y_i) = \pi(\underline{x}_i, \underline{\beta})\Big(1 - \pi(\underline{x}_i, \underline{\beta})\Big). \tag{2}$$

It is also commonly assumed that the responses are independent across the observations. Under these assumptions, the likelihood function with respect to $\underline{\beta}$ can be obtained from the joint probability mass function

*University of Wyoming 1000 E University Ave, Laramie, WY 82071; Covance Inc. 1016 West Ninth Ave., King of Prussia, PA 19406

†University of Wyoming, 1000 E University Ave, Laramie, WY 82071

‡University of Wyoming, 1000 E University Ave, Laramie, WY 82071

$$L = L(\underline{\beta}|\underline{y}) = \prod_{i=1}^{N} \pi(\underline{x}_i, \underline{\beta})^{y_i}(1 - \pi(\underline{x}_i, \underline{\beta}))^{1-y_i}. \tag{3}$$

The estimated probability of a positive outcome ($\hat{\pi}_i$) can be obtained using the maximum likelihood estimator (MLE) ($\hat{\underline{\beta}}$) by finding the value of $\underline{\beta}$ that maximizes (3). As a result, the estimated probability of positive outcome for observation $i$ is $\hat{\pi}_i = \pi(\underline{x}_i, \hat{\underline{\beta}}), i = 1, 2, ...N$.

### 1.2 Link functions

The *Bernoulli* distribution is a member of the exponential family of distributions, and thus can be modeled in the generalized linear model (GLM) framework. The *link function* is a one-to-one continuous differentiable transformation of the expected value of the random variable ($g(\mu_i)$). The *linear predictor* is defined as $\eta_i = \eta(\underline{x}_i, \underline{\beta}) = \underline{x}_i'\underline{\beta}$, so that

$$g(\mu_i) = \eta_i = \eta(\underline{x}_i, \underline{\beta}) = \underline{x}_i'\underline{\beta}. \tag{4}$$

Using the inverse function $g^{-1}(\cdot)$, the above equation can be expressed as

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\Big(\underline{x}_i'\underline{\beta}\Big). \tag{5}$$

Different link functions have been applied in logistic regression, including the identity link, the logit link, the probit link, and the complementary log-log link (Agresti (1990) pp. 86, 103, 105). The logit link

$$g(\mu_i) = logit\Big(\pi(\underline{x}_i, \underline{\beta})\Big) = \log\Big(\frac{\pi(\underline{x}_i, \underline{\beta})}{1 - \pi(\underline{x}_i, \underline{\beta})}\Big) = \underline{x}_i'\underline{\beta}, \tag{6}$$

is the canonical link and it has a fairly simple interpretation as the log of odds of positive outcome or $\log\Big(\frac{\pi(\underline{x}_i, \underline{\beta})}{1-\pi(\underline{x}_i, \underline{\beta})}\Big)$.

## 2. Model selection for logistic regression

Model selection refers to the process of selecting a 'best-fit' model among several competing candidate models. Two models are *nested* if one of the models (the *reduced* model) can be obtained by setting some of the regression coefficients equal to zero in the other model (the *full* model).

### 2.1 Deviance and likelihood ratio test

A likelihood ratio test can be conducted by comparing a reduced model to a full model in terms of *deviance*, which is the negative two times the maximum log likelihood, i.e., $D = -2\ln(\hat{L})$, where $\hat{L} = L(\hat{\underline{\beta}})$. Under the null hypothesis that the reduced model is the true model, the difference between deviances of the full model and the reduced model asymptotically follows a chi-square distribution (Myers et al. (2010)), with the degrees of freedom are equal to the difference ($\Delta$) in the number of regression coefficients between the nested models,

$$D(reduced) - D(full) = -2\ln\Big[\frac{L(\hat{\underline{\beta}}_{reduced})}{L(\hat{\underline{\beta}}_{full})}\Big] = -2\ln(\Lambda) \sim \chi^2_{\Delta}. \tag{7}$$

### 2.2 AIC and BIC

The Akaike information criterion (AIC) is another tool for model selection

$$AIC = -2\ln(\hat{L}) + 2p, \tag{8}$$

where $\hat{L}$ is the maximized value of the likelihood function for the model, and $p$ is the number of model parameters (Akaike (1973) and Akaike (1974)). The preferred model is the one with the minimum AIC value over the candidate set of models.

The Bayesian information criterion (BIC) is defined as (Schwarz (1978))

$$BIC = -2\ln(\hat{L}) + p \times ln(N), \tag{9}$$

where the model with the lowest BIC among those in the candidate set is preferred.

### 2.3 Sensitivity, specificity, and ROC

Sensitivity and specificity are often used to measure the predictive performance of the model fit. A *true positive* means that the observation was predicted to be positive and it was observed to be positive. A *false positive* means that the observation was predicted to be positive but was observed to be negative. A *true negative* means that the observation was predicted to be negative and it was observed to be negative. A *false negative* means that the observation was predicted to be negative but was observed to be positive. Accordingly, *sensitivity* refers to the ability of the fitted model to correctly identify the positive observations as given by

$$Sensitivity = \frac{\#\ of\ true\ positives}{\#\ of\ true\ positive + \#\ of\ false\ negatives}, \tag{10}$$

while *specificity* refers to the ability of the fitted model to correctly identify the negative observations as given by

$$Specificity = \frac{\#\ of\ true\ negatives}{\#\ of\ true\ negatives + \#\ of\ false\ positives}. \tag{11}$$

The estimated probabilities from logistic regression are used for classification of the positive or negative outcomes. A predicted classification of 'positive' occurs if the value of the estimated probability ($\hat{\pi}_i$) is large while a predicted outcome of 'negative' occurs if the estimated probability ($\hat{\pi}_i$) is small. A cut-point ($\pi^c$) is used to define a classification rule quantifying large and small in which $\hat{y}_i = 1$ if $\hat{\pi}_i > \pi^c$ (predicted positive) and $\hat{y}_i = 0$ (predicted negative) if $\hat{\pi}_i \leq \pi^c$ (Kutner et al. (2005) pp. 604-605). The $sensitivity$ and $specificity$ depend upon the cut-point value ($\pi^c$). A complete description of the classification ability of a model is given by the Receiver Operating Characteristic (ROC) curve (Hosmer et al. (2013), Section 5.2.4), with the ROC curve being a plot of the $sensitivity$ versus $1-specificity$ across a range of potential cut-point values. The area under the ROC curve provides a measure of the predictive ability of the model where larger areas suggest better classification than smaller areas. Hosmer et al. (2013) (p.177) give the following general guidelines:

$$0.5 \leq area\ under\ ROC < 0.7 \implies poor\ classification, \tag{12}$$

$$0.7 \leq area\ under\ ROC < 0.8 \implies acceptable\ classification, \tag{13}$$

$$0.8 \leq area\ under\ ROC < 0.9 \implies excellent\ classification, \tag{14}$$

$$0.9 \leq area\ under\ ROC \leq 1.0 \implies outstanding\ classification. \tag{15}$$

## 3. Goodness-of-fit for logistic regression

Goodness-of-fit tests involve the null hypothesis of whether the modeled distribution $F$ for the responses satisfies $H_0 : F = F_0$ where $F_0$ is a specified distribution (Lehmann (1998)). A fitted model can be inadequate because the linear systematic component of the model may be incorrectly specified, a covariate may not be specified in the appropriate functional form, some important covariates may have been omitted from the model, or the link function may be misspecified (Xie et al. (2008)). All these could affect the consistency of the coefficient estimation, and lead to biased estimates of treatment effects (Gail et al. (1988); Hauck et al. (1991)).

Many goodness-of-fit tests have been proposed during the past four decades. The following sections focus on chi-quare and deviance based tests, goodness-of-fit tests based upon residuals, and other types of goodness-of-fit tests.

### 3.1 Chi-square and deviance based tests

Chi-square and deviance based goodness-of-fit tests depend on the type of covariates in the model and the number of covariate patterns of regressors. The covariate pattern represents a single set of values for the covariates in a model (Hosmer and Lemeshow (2000)). The following paragraphs summarize three types of covariates and the goodness-of-fit tests that can be applied.

(1) All the covariates in a logistic regression model are categorical regressors and the number of covariate patterns is small relative to the number of positive responses. The Pearson chi-square test and the deviance test are typically relevant for assessing model fit.

(2) The number of covariate patterns is large relative to the number of positive responses. This happens when there are continuous covariates in the logistic model, or the sample size is relatively small compared to the number of covariate patterns. The Hosmer and Lemeshow (1980) approach can be considered for assessing goodness-of-fit.

(3) Both continuous and categorical covariates exist. This scenario can be considered to be a special case of (2). The Pulkstenis and Robinson (2002) and the Xie et al. (2008) tests can be considered for assessing goodness-of-fit.

#### *3.1.1 Pearson chi-square test and deviance test*

Consider $N$ outcomes from $G$ groups, with $N = \sum_{g=1}^{G} n_g$, where $n_g$ represents the number of observations in group $g$. Let a grouped binary response $Y_g$ denote the number of positive outcomes out of $n_g$ observations and $\underline{x}_g = (1, x_{1g}, ...x_{cg})'$ denote the $c$ categorical regressors in group $g$. Let $o_{g,1}$ represent the number observed positive outcomes, and $o_{g,0}$ represent the number observed negative outcomes in group $g$. The expected number of positive outcomes can be calculated as $e_{g,1} = n_g \times \hat{\pi}(\underline{x}_g, \underline{\hat{\beta}})$, and the expected number of negative outcomes can be expressed as $e_{g,0} = n_g - e_{g,1}$.

The Pearson chi-square statistic (Agresti (1990)) can then be calculated as

$$\hat{C}_P = \sum_{g=1}^{G} \frac{(o_{g,1} - e_{g,1})^2}{e_{g,1}} + \sum_{g=1}^{G} \frac{(o_{g,0} - e_{g,0})^2}{e_{g,0}}. \tag{16}$$

The deviance statistic from the likelihood ratio test (Myers et al. (2010)) can be expressed as

$$D^2_{df} = -2 \sum_{g=1}^{G} \Big[ o_{g,1} \log \big(\frac{o_{g,1}}{e_{g,1}}\big) + o_{g,0} \log \big(\frac{o_{g,0}}{e_{g,0}}\big) \Big]. \tag{17}$$

Under the null hypothesis that the fitted model is correctly specified, $\hat{C}_P$ and $D^2_{df}$ are asymptotically chi-square distributed with degrees of freedom $df = G - (c + 1)$ (Agresti (1990)), with $c$ being the number of regressors in the model. There are many instances in which these two types of statistics give similar results (Myers et al. (2010)).

### *3.1.2 Hosmer-Lemeshow test*

When all regressors in a logistic regression are continuous covariates, Hosmer and Lemeshow (1980) propose an approach to group the subjects based on estimated probabilities, and then compute a chi-square statistic. Two grouping methods have been proposed. One method is to collapse the table based on percentiles of the estimated probabilities. The other method is to collapse the table based on fixed values of the estimated probability. For either grouping strategy, the Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating a chi-square statistic from the $G \times 2$ table as

$$\hat{C}_{HL} = \sum_{j=0}^{1} \sum_{g=1}^{G} \frac{(o_{g,j} - e_{g,j})^2}{e_{g,j}}, \tag{18}$$

where $o_{g,j}$ denotes the observed number of positive (when $j = 1$) or negative outcomes (when $j = 0$) in group $g$, and $e_{g,j}$ denotes the estimated number of positive (when $j = 1$) or negative outcomes (when $j = 0$) in the group $g$. A chi-square test with degrees of freedom $df = G - 2$ on the statistic ($\hat{C}_{HL}$) has been most commonly used in practice (Hosmer and Lemeshow (2000)). Further research by Hosmer et al. (1988) has indicated that the first grouping method is preferable to the second based on fixed cutpoints.

### *3.1.3 Pulkstenis and Robinson tests*

When both continuous and categorical regressors exist, Pulkstenis and Robinson (2002) propose a two-level subgrouping based on fitted probabilities within each covariate pattern. This approach requires sorting all responses by fitted probabilities within each unique covariate pattern as defined only by categorical regressors, and then creating two subcategories by splitting the responses based on the median of fitted probabilities within each covariate pattern. The proposed test statistics are given by the following modified chi-square ($\hat{C}_{PR}$) and deviance ($D^2_{PR}$) statistics:

$$\hat{C}_{PR} = \sum_{m=1}^{M} \sum_{h=1}^{2} \sum_{j=0}^{1} \frac{(o_{m,h,j} - e_{m,h,j})^2}{e_{m,h,j}}, \tag{19}$$

$$D^2_{PR} = -2 \sum_{m=1}^{M} \sum_{h=1}^{2} \sum_{j=0}^{1} o_{m,h,j} \log \frac{o_{m,h,j}}{e_{m,h,j}}, \tag{20}$$

where $m = 1, \ldots, M$ indexes covariate patterns based on categorical regressors, $h = 1, 2$ indexes the substratification due to the median split of the fitted probabilities within each covariate pattern, and $j = 0, 1$ indexes the two response categories (negative or positive outcome). Through simulation studies, Pulkstenis and Robinson (2002) suggest a chi-square test can be conducted on the modified chi-square ($\hat{C}_{PR}$) and deviance ($D^2_{PR}$) statistics with $2M - c - 2$ degrees of freedom, where $c$ is the number of variables in the model needed to represent all non-continuous covariates.

### 3.1.4 *Tsiatis score test*

Tsiatis (1980) proposes a goodness-of-fit test, where the space of covariates $(Z_1, \ldots, Z_m)$ is partitioned into $k$ distinct regions in $m$-dimensional space denoted by $R_1, \ldots, R_k$. The indicator function is defined by $I^j (j = 1, \ldots, k)$ in which $I^j = 1$ if $(Z_1, \ldots, Z_m \in R_j)$ and $I^j = 0$ otherwise. Consider the following logistic regression model

$$\log(\frac{p_z}{1-p_z}) = \underline{\beta}^{'} \underline{Z} + \underline{\gamma}^{'} \underline{I}, \tag{21}$$

where $\underline{I}^{'} = (I^1, \ldots, I^k)$ is the indicator function and $\underline{\gamma}^{'} = (\gamma_1, \ldots, \gamma_k)$ denotes the shift parameters. Here $\underline{\beta}^{'} \underline{Z}$ accounts for all of the original covariates and $\underline{\gamma}^{'} \underline{I}$ accounts for the regional shifts. In order to test the null hypothesis $H_0 : \gamma_1 = \cdots = \gamma_k = 0$ against the alternative hypothesis $H_1$ : at least one $\gamma_i \neq 0$, Tsiatis (1980) forms a test statistic based on the efficient scores test

$$T = X^{'} V^{-} X, \tag{22}$$

where $X$ denotes the $k$-dimensional vector $(\partial l/\partial\gamma_1, \ldots, \partial l/\partial\gamma_k)$, $l$ denotes the log-likelihood, and $V^{-}$ denotes the g-inverse of $V$. The $k \times k$ matrix $V$ is equal to

$$V = A - BC^{-1}B^{'}, \tag{23}$$

where $A_{jj'} = -\partial^2 l/\partial\gamma_j\partial\gamma_{j'}$ $(j, j^{'} = 1, ...k), B_{jj'} = -\partial^2 l/\partial\gamma_j\partial\beta_{j'}$ $(j = 1, ...k; j^{'} = 0, ..., m)$, and $C_{jj'} = -\partial^2 l/\partial\beta_j\partial\beta_{j'}$ $(j, j^{'} = 0, ..., m)$. The above terms are evaluated at $\underline{\gamma} = \underline{0}$ and $\underline{\beta} = \underline{\hat{\beta}}$, where $\underline{\hat{\beta}}$ is the maximum likelihood estimate of the parameters when $H_0$ is true. Under the null hypothesis, the statistic $T$ is asymptotically distributed as a chi-squared distribution with degrees of freedom equal to the rank of $V$. While Tsiatis approach is conceptually elegant, it lacks a general rule for how to partition the covariate space, especially when continuous covariates are present (Xie et al. (2008)).

### 3.1.5 *Xie tests*

Xie et al. (2008) integrate cluster analysis, the Pearson chi-square test, and the Tsiatis (1980) score test together to form a chi-square test and a score test. The tests contain two steps. First conduct a cluster analysis on all covariates to group observations into $G$ clusters. Second, calculate a chi-square statistic or score statistic on these clusters.

The Xie chi-square test ( $\hat{C}_{Xie}$) involves constructing the chi-square statistic,

$$\hat{C}_{Xie} = \sum_{g=1}^{G} \frac{(o_{g,1} - n_g^{'}\bar{\pi}_g)^2}{n_g^{'}\bar{\pi}_g(1-\bar{\pi}_g)}, \tag{24}$$

where $n_g^{'}$ denotes the number of observations in cluster $g$, $o_{g,1}$ denotes the observed number of positive outcomes in cluster $g$, and $\bar{\pi}_g$ denotes the average estimated probability for all observations in cluster $g$. Based on simulation studies, Xie et al. (2008) propose using $G = 10$ if $k < 5$ and $G = k + 5$ if $k \geq 5$, then comparing $\hat{C}_{Xie}$ to an asymptotic chi-square distribution with degrees of freedom $df = G - (k/2) - 1$, with $k$ being the number of covariates in the model.

By adding a shift parameter $\underline{\gamma}$ in each region to the Tsiatis (1980) test via a series of indicator functions, Xie et al. (2008) propose a modified score test. The extended logistic regression model is gives by

$$\log(\frac{\pi_i}{1-\pi_i}) = \underline{x_i}^{'}\underline{\beta} + \underline{\gamma}^{'} \underline{I}_i, \tag{25}$$

where $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^{'}$, $\underline{x}_i = (1, x_{1i}, x_{2i}, \ldots, x_{ki})^{'}$, $\underline{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{G-1})^{'}$, and $\underline{I}_i = (I_i^{(1)}, I_i^{(2)}, \ldots, I_i^{(G-1)})$. The goodness-of-fit test consists of testing the hypothesis $H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_{G-1} = 0$ against the alternative $H_A$: at least one $\gamma_g \neq 0$, for $g = 1, ..., G-1$. The Xie score test is based on the efficient score statistic $T = U^{'}V^{-}U$, where $U$ denotes the $G-1$ dimensional vector $(\partial l/\partial\gamma_1, \ldots, \partial l/\partial\gamma_{G-1})$, and $l$ denotes the log-likelihood. The $(G-1) \times (G-1)$ matrix $V$ can be expressed as $V = A - BC^{-1}B^{'}$, where $A_{jj'} = -\partial^2 l/\partial\gamma_j\partial\gamma_{j'}(j, j^{'} = 1, \ldots, G-1)$, $B_{jj'} = -\partial^2 l/\partial\gamma_j\partial\beta_{j'}(j = 1, \ldots, G-1; j^{'} = 0, 1, \ldots k)$, $C_{jj'} = -\partial^2 l/\partial\beta_j\partial\beta_{j'}(j, j' = 0, 1, \ldots, k)$ (Rao (1973), Tsiatis (1980)). All the above terms are evaluated at $\underline{\gamma} = \underline{0}$ and $\underline{\beta} = \underline{\hat{\beta}}$, with $\underline{\hat{\beta}}$ being the maximum likelihood estimate of $\underline{\beta}$ under $H_0$. Under the null hypothesis, $T$ is asymptotically distributed as a chi-square distribution with degrees of freedom at rank($V$).

### 3.2 Goodness-of-fit based on residuals

In standard linear regression, it is common to use residuals for model diagnostics and to assess goodness-of-fit. For other generalized linear models (e.g., logistic regression), the choice of residuals may not be obvious nor how to use them to assess the model fit. The next few subsections discuss generalized residuals which have been proposed for model diagnosis in logistic regression, as well as goodness-of-fit tests that are based upon such residuals.

#### *3.2.1 Pearson residual and deviance residual*

(1) The Pearson residual is given by

$$r_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}. \tag{26}$$

(2) The deviance residual is given by

$$d_i = sgn(y_i - n_i\hat{\pi}_i)\Big\{2\Big[y_i \log\Big(\frac{y_i}{n_i\hat{\pi}_i}\Big) + (n_i - y_i)\log\Big(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\Big)\Big]\Big\}^{1/2}. \tag{27}$$

The Pearson residuals and deviance residuals correspond to the Pearson chi-square test and the deviance test.

#### *3.2.2 Nonparametric kernel method*

Nonparametric methods can also been applied to examine the residuals in logistic regression. Consider a single regressor where $g(x) = Pr(Y = y|X = x)$. Then the associated hypothesized logistic model with a single predictor, $g_0(x)$, is given by

$$logit(g_0(x)) = \beta_0 + \beta_1 x. \tag{28}$$

Copas (1983) introduces a nonparametric kernel method to examine the model-fit graphically. This approach has been extended by Azzalini et al. (1989), who propose an approach to compare the function $g_0(x)$ with kernel estimate $\tilde{g}(x)$ of $g(x)$. Through simulation, the confidence bands for the nonparameteric curve can be obtained under the null hypothesis. Azzalini et al. (1989) generate a pseudo-likelihood ratio statistic by comparing the function $g_0(x)$ with $\tilde{g}(x)$.

le Cessie and van Houwelingen (1991) further refine the approach of Azzalini et al. (1989) using an unbiased estimator. The residuals of the model, $y_i - g_0(x_i)$, can be standardized under the null hypothesis as

$$r_i = \frac{y_i - g_0(x_i)}{\sqrt{g_0(x_i)(1 - g_0(x_i))}}. \tag{29}$$

A smoothing function of these standardized residuals is defined by

$$\tilde{r}_i = \frac{\sum_j r_j K[(x_i - x_j)/h_n]}{\sum_j K[(x_i - x_j)/h_n]}, \tag{30}$$

where $h_n$ is the bandwidth that controls the amount of smoothing. The function $K$ is a nonnegative symmetric bounded kernel function, zero outside a closed interval $[-a, a]$, and is normalized according to $\int K(z)dz = 1$ and $\int K(z)^2 dz = 1$.

A weighted sum of the smoothed standardized residuals is used as the goodness-of-fit measure. The test statistic $T$ is defined by

$$T = \frac{1}{n}\sum_{i=1}^{n} \tilde{r}_i^2 v_i, \tag{31}$$

where $v_i = \frac{\{\sum_j K[(x_i - x_j)/h_n]\}^2}{\sum_j K[(x_i - x_j)/h_n]^2}$ is the inverse of the variance of the smoothed standardized residual.

Under the null hypothesis, $g(x)$ is close to $g_0(x)$, and the expected value of $T$ conditional on the predictor $x$ is 1. The variance of $T$ has complex form but can be calculated exactly as

$$var(T) = n^{-2}\sum_i\sum_j \Big[\sum_k w_{ik}^2 w_{jk}^2\Big]^{-1}\Big[\sum_k \frac{w_{ik}^2 w_{jk}^2(6g_0(x_k)^2 - 6g_0(x_k) + 1)}{g_0(x_k)(1 - g_0(x_k))} + 2\Big(\sum_k w_{ik}w_{jk}\Big)^2\Big], \tag{32}$$

where $w_{ij} = K[(x_i - x_j)/h_n]$. Consequently, the test statistic and associated p-value can be calculated as in the le Cessie and van Houwelingen (1991) test.

This approach can be extended to higher dimensions with multiple covariates. The choice of the bandwidth is crucial, which depends on the number of observations, the number of variables, and the alternative hypotheses. le Cessie and van Houwelingen (1991) suggest a bandwidth such that each region over which the residuals are averaged contains approximately $\sqrt{n}$ observations.

### *3.2.3 Generalized $R^2$ coefficients*

Generalized $R^2$ coefficients have been developed for logistic regression. Mittlbock and Schemper (1996) study the properties of 12 different $R^2$ measures and recommend two $R^2$ coefficients for routine use.

(1) The squared Pearson correlation coefficient of observed outcomes with the predicted probabilities is defined as follows

$$r^2 = \frac{[\sum_{i=1}^n (y_i - \bar{y})(\hat{\pi}_i - \bar{\pi})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{\pi}_i - \bar{\pi})^2}, \tag{33}$$

with $n$ denotes the number of covariate patterns.

(2) The linear regression-like sum-of-square $R^2$, which is defined as

$$R^2_{ss} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}. \tag{34}$$

Mittlbock and Schemper (1996) also recommend

$$R^2_l = \frac{l_0 - l_p}{l_0} = 1 - \frac{l_p}{l_0}, \tag{35}$$

where $l_0$ and $l_p$ denote the log-likelihoods for the models containing only the intercept versus the model containing the intercept plus $p$ covariates, respectively. However, Hosmer and Lemeshow (2000) suggest these $R^2$ coefficients typically have low values in logistic regression and are not good measure for assessing model fit.

### 3.3 Other goodness-of-fit tests

There are a number of goodness-of-fit tests in addition to those previously described. White (1982) proposes a test to detect model mis-specification based on information matrix. Newey (1985) further discusses the information matrix test for probit models and notes that this approach is sensitive to heteroskedasticity and non-normality. He proposes a simple calculation procedure which employs the "outer product of the gradient" (OPG) covariance matrix estimator of the information matrix test statistic. Orme (1988) proposes a simple calculation procedure for the information matrix test statistic for general binary data models by employing the maximum likelihood covariance matrix estimator rather than the OPG estimator. Stukel (1988) incoporates two shape parameters to extend the formulation of logistic model and improve the model fit. Hosmer et al. (1997) suggest that the Stukel test has higher power than other tests they examined for misspecified link functions and comparable power to other tests overall. Copas (1989) conducts a study on the unweighted residual sum of squares, which can be considered as a modification to the original Pearson chi-square statistics. However, according to Copas (1989), a major disadvantage is that the test statistic no longer has a chi-square distribution, even asymptotically. Further studies are still needed to examine whether the scaled chi-square distribution is a good approximation on the null distribution of the test statistic. The unweighted sum-of-squares goodness-of-fit test is described in Hosmer et al. (1997) and is frequently termed the "le Cessie-van Houwelingen-Copas-Hosmer goodness of fit test". Hosmer et al. (1997) find this test to be superior to the other tests they examined based on overall performance. McCullagh (1985) proposes a goodness-of-fit test for logistic regression models using conditional asymptotic moments from Pearson chi-square statistics. Farrington (1996) extend McCullagh (1985) test based on conditioning principles. He proposes a test using first-order modification to the Pearson statistic. Osius and Rojek (1992) derive asymptotic moments for a general class of goodness-of-fit statistics (power-divergence family), and use that to conduct a standardized test statistic. The p-value is calculated by comparing the statistic to a standard normal distribution. Qin and Zhang (1997) test the logistic regression assumption under a case-control sampling plan, and propose a Kolmogorov-Smirnov statistic to test the validity of the logistic link function.

## 4. Conclusion

This paper provides an overview of logistic regression. Several model selection procedures were described, including the likelihood-ratio test, AIC and BIC, sensitivity, specificity and the ROC curve. Goodness-of-fit tests were discussed with emphasis on chi-square and deviance approaches as well as nonparametric methods. When only categorical covariates

exist, one can apply the Pearson chi-square test and the deviance test. When continuous regressors exist in a logistic regression, one can apply the Hosmer and Lemeshow (1980) test. When both continuous and categorical covariates exist in the model, one can apply the Pulkstenis and Robinson (2002) tests and the Xie et al. (2008) tests in conjunction with the standard Hosmer and Lemeshow (1980) test. The nonparametric kernel methods developed by le Cessie and van Houwelingen (1991) can also be applied to logistic regressions with multiple covariates, but the test results could be sensitive to the choice of the bandwidth from the smoothing functions.

This overview describes many of the important criteria for assessing logistic regression models. Further research could involve the incorporation of all of these criteria into the model selection process. Multiobjective decision making tools may be useful for identifying a suitable logistic regression model based upon these multiple conflicting criteria.

## 5. Appendix: R functions for model assessment in logistic regression

**Logistic regression**
**glm (formula, family = binomial** ***(link='logit', 'probit', or 'cloglog')*** ) can be used for logistic regression with link function chosen from the logit link, the probit link, and the complementary log-log link (stats package). **fitted** is a generic function which extracts fitted values from objects returned by modeling functions. **fitted.values** is an alias for it.

**Model selection**
**logLik:** is a generic function to calculate the log-likelihood of a model (stats package). The model deviance can be calculated by **-2 * logLik**.
**pchisq:** is the distribution function for the chi-squared distribution with df degrees of freedom (stats package). It can be used to compute the p-value for chi-square based tests.
**AIC:** is a generic function to calculate the Akaike information criterion (stats package).
**BIC:** is a generic function to calculate the Bayesian information criterion (stats package).
**roc.area:** is a function to calculate the area uderneath a ROC curve (verification package).

**Goodness-of-fit**
**Pearson chi-square test statistic** can be calculated by **sum(((tapply(y, g, sum) - tapply(phat, g, sum))^2) / tapply(phat, g, sum)) + sum (((tapply((1-y), g, sum) - tapply((1-phat), g, sum))^2) / tapply((1-phat), g, sum))**, and the **deviance test** can be calculated by **sum(2*tapply(y, g, sum) * log(tapply(y, g, sum) / tapply(phat, g, sum))) + sum (2 * tapply((1-y), g, sum) * log(tapply((1-y), g, sum) / tapply((1-phat), g, sum)))**, where y is the observed value (0 or 1), phat is the fitted value from logistic regression model, and $g$ is the group factor for each observation.
**logitgof:** performs the Hosmer-Lemeshow goodness-of-fit test for **binary, multinomial** and **ordinal** logistic regression models (generalhoslem package).
**pulkrob.chis, pulkrob.deviance:** perform the Pulkstenis and Robinson (2004) chi-square and deviance tests for **ordinal** logistic regression models (generalhoslem package). Please contact the authors for sample R codes for the Pulkstenis and Robinson (2002) chi-square and deviance tests on **binary** responses.
**PseudoR2 (mod, 'all'):** is a function to compute several variants of pseudo $R^2$ statistics for logistic regression model, including the AldrichNelson pseudo-$R^2$, the McFadden pseudo-$R^2$, the McFadden adjusted pseudo-$R^2$, the Cox and Snell pseudo-$R^2$, the Nagelkerke pseudo-$R^2$, the McKelvey and Zavoina pseudo-$R^2$, the Effron pseudo-$R^2$, and the Tjur pseudo-$R^2$ (DescTools package).

**1 - logLik(current model) / logLik(null model):** can be used to compute the Mittlbock and Schemper $R^2$ based on the log-likelihoods for the null model (only containing the intercept) versus current model (containing the intercept plus $k$ covariates), respectively. **resid** is the function in the Design package for the le Cessie and Houwelingen test, though it requires using the **lrm** function for logistic regression. Package 'Design' was removed from the CRAN repository. Formerly available versions can be obtained from the archive (https://cran.r-project.org/src/contrib/Archive/Design/), and can be installed to earlier version of R.

## References


A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 1990.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory*, pages 267–281, Budapest: Akademia Kiado, 1973.

H. Akaike. A new look at the statistical model identification. In *IEEE Transactions on Automatic Control*, pages 716–723, 1974.

A. Azzalini, A. Bowman, and A. W. Hardle. On the use of nonparametric regression for model checking. *Biometrika*, 76(1):1–11, 1989.

J. B. Copas. Plotting p against x. *Appl. Statist.*, 32:25–31, 1983.

J. B. Copas. Unweighted sum of squares test for proportions. *Appl. Statist.*, 38(1):71–80, 1989.

C. P. Farrington. On assessing goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society, Series B*, 58(2):349–360, 1996.

M. H. Gail, W.Y. Tan, and S. Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 75(1):57–64, 1988.

W. W. Hauck, J. M. Neuhaus, J. D. Kalbfleisch, and S. Anderson. A consequence of omitted covariates when estimating odds ratios. *J. Clin. Epidemiol.*, 44(1):77–81, 1991.

D. W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Commun. Statist. PartA-Theory and Methods*, pages 1043–69, 1980.

D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2th edition, 2000.

D. W. Hosmer, S. Lemeshow, and J. Klar. Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biometrical Journal*, 30(8):911–924, 1988.

D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statist. Med.*, 16:956–980, 1997.

D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 3th edition, 2013.

M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill: Boston, 2005.

S. le Cessie and C. van Houwelingen. A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrika*, 47:1267–82, 1991.

E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, 2th edition, 1998.

P. McCullagh. On the asymptotic distribution of pearson's statistic in linear exponential-family models. *International Statistical Review*, 53(1):61–67, 1985.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2th edition, 1989.

M. Mittlbock and M. Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15:1987–97, 1996.

R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson. *Generalized linear models with applications in engineering and the sciences*. John Wiley & Sons, 2th edition, 2010.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, pages 370–384, 1972.

W. K. Newey. Maximum likelihood specification testing and conditional moment tests. *Econometrica*, 53:1047–70, 1985.

C. Orme. The calculation of the information matrix test for binary data models. *The Manchester School*, 54(4):370–376, 1988.

G. Osius and D. Rojek. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, 87(420):1145–52, 1992.

E. Pulkstenis and T. J. Robinson. Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statist. Med.*, 21:79–93, 2002.

E. Pulkstenis and T. J. Robinson. Goodness-of-fit tests for ordinal response regression models. *Statist. Med.*, 23:999–1014, 2004.

J. Qin and B. Zhang. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618, 1997.

C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, 2th edition, 1973.

E. G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

T. A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83:426–431, 1988.

A. A. Tsiatis. A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, 67:250–251, 1980.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1): 1–25, 1982.

X. J. Xie, J. Pendergast, and W. Clarke. Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with coninuous predictors. *Computational Statistics & Data Analysis*, 52:2703–13, 2008.