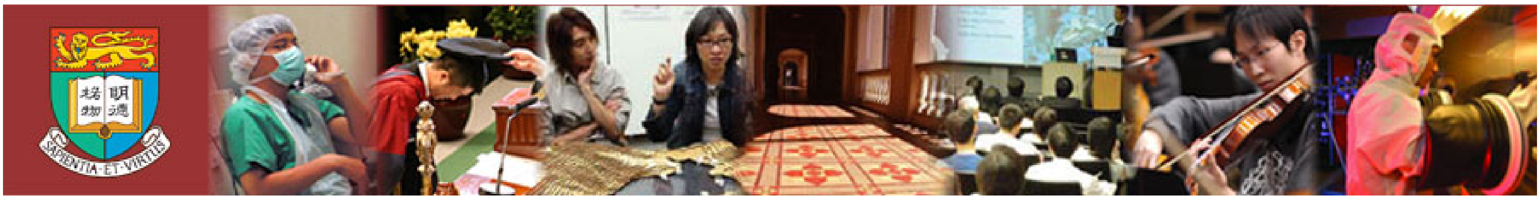


Title	Pearson-type goodness-of-fit test with bootstrap maximum likelihood estimation
Author(s)	Yin, G; Ma, Y
Citation	Electronic Journal of Statistics, 2013, v. 7, p. 412-427
Issued Date	2013
URL	http://hdl.handle.net/10722/189457
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Electronic Journal of Statistics
Vol. 7 (2013) 412–427
ISSN: 1935-7524
DOI: 10.1214/13-EJS773

Pearson-type goodness-of-fit test with bootstrap maximum likelihood estimation

Guosheng Yin

Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road, Hong Kong
e-mail: gyin@hku.hk

and

Yanyuan Ma

Department of Statistics
Texas A&M University
College Station, Texas 77843, U.S.A.
e-mail: ma@stat.tamu.edu

Abstract: The Pearson test statistic is constructed by partitioning the data into bins and computing the difference between the observed and expected counts in these bins. If the maximum likelihood estimator (MLE) of the original data is used, the statistic generally does not follow a chi-squared distribution or any explicit distribution. We propose a bootstrap-based modification of the Pearson test statistic to recover the chi-squared distribution. We compute the observed and expected counts in the partitioned bins by using the MLE obtained from a bootstrap sample. This bootstrap-sample MLE adjusts exactly the right amount of randomness to the test statistic, and recovers the chi-squared distribution. The bootstrap chi-squared test is easy to implement, as it only requires fitting exactly the same model to the bootstrap data to obtain the corresponding MLE, and then constructs the bin counts based on the original data. We examine the test size and power of the new model diagnostic procedure using simulation studies and illustrate it with a real data set.

AMS 2000 subject classifications: Primary 62J20, 62F40; secondary 62J12.
Keywords and phrases: Asymptotic distribution, bootstrap sample, hypothesis testing, maximum likelihood estimator, model diagnostics.

Received June 2012.

Contents

1. Introduction

The model goodness-of-fit test is an important component of model fitting, because model misspecification may cause severe bias and even lead to incorrect inference. The classical Pearson chi-squared test can be traced back to the pioneering work of Pearson (1900). Since then, various model selection and diagnostic tests have been proposed in the literature (Claeskens and Hjort, 2008). In contrast to model selection which concerns multiple models under consideration and eventually selects the best fitting model among them, model diagnostic tests are constructed for a single model, and the goal is to examine whether the model fits the data adequately. The commonly used criterion-based model selection procedures include the Akaike information criterion (AIC) by Akaike (1973) and Bayesian information criterion (BIC), which, however, cannot be used for testing the fit of a single model. For model diagnostics, a common practice is to plot the model residuals versus the predictive outcomes. If the model fits the data adequately, we expect the residuals would be fluctuating around the zero axis, which can thus be used as a graphical checking tool for model misspecification. More sophisticated statistical tests may be constructed based on the partial or cumulative sum of residuals (for example, see Su and Wei, 1991; Stute, Manteiga and Quindimilm, 1998; and Stute and Zhu, 2002).

The classical Pearson chi-squared test statistic is constructed by computing the expected and observed counts in the partitioned bins (Pearson, 1900). More specifically, let $(y_1, \dots, y_n)$ denote a random sample from the distribution $F_{\beta_0}(y)$, where β_0 is the true parameter characterizing the distribution function. We are interested in examining whether the sample is from $F_{\beta_0}(y)$; that is, the null hypothesis is $H_0 : F_{\beta_0}(y)$ is the true distribution for the observed data, and the alternative is $H_1 : F_{\beta_0}(y)$ is not the true distribution for the observed data. In the Pearson test, we first partition the sample space into K nonoverlapping bins, and let p_k denote the probability assigned to bin k, for $k = 1, \dots, K$. When the true parameter value β_0 is known, we can easily count the number of observations falling into each prespecified bin. We denote the observed count for bin k as m_k. The Pearson goodness-of-fit test statistic takes the form of

$$Q_1(\beta_0) = \sum_{k=1}^{K} \frac{(m_k - np_k)^2}{np_k}, \tag{1.1}$$

which asymptotically follows the $\chi^2_{(K-1)}$ distribution under the null hypothesis. We may replace the expected counts np_k in the denominator of (1.1) by the observed counts m_k,

$$Q_2(\beta_0) = \sum_{k=1}^{K} \frac{(m_k - np_k)^2}{m_k}, \tag{1.2}$$

which is asymptotically equivalent to $Q_1(\beta_0)$, and also follows the $\chi^2_{(K-1)}$ distribution.

However, the true parameter β_0 is often unknown in practice. As a consequence, we need to estimate β_0 in order to construct the bin probabilities or bin counts. For non-regression settings with independent and identically distributed (i.i.d.) data, Chernoff and Lehmann (1954) showed that using the maximum likelihood estimator (MLE) of β_0 based on the original data, $\hat{\beta}$, the test statistic does not follow a χ^2 distribution or any explicit known distribution. In particular, we denote the corresponding estimates for the bin probabilities by $p_k(\hat{\beta})$, and define

$$Q(\hat{\beta}) = \sum_{k=1}^{K} \frac{\{m_k - np_k(\hat{\beta})\}^2}{np_k(\hat{\beta})}.$$

Generally speaking, $Q(\hat{\beta})$ does not follow a χ^2 distribution asymptotically, but it stochastically lies between two χ^2 distributions with different degrees of freedom. This feature of the Pearson-type χ^2 test weakens its generality and limits its applicability to a variety of regression models for which the maximum likelihood estimation procedure dominates. Although some numerical procedures can be used to approximate the null distribution, but they are typically quite computationally intensive (e.g., see Imhof, 1961; and Ali, 1984). If we apply the maximum likelihood estimation to the grouped data and denote the corresponding MLE as $\hat{\beta}_g$, then

$$Q(\hat{\beta}_g) = \sum_{k=1}^{K} \frac{\{m_k - np_k(\hat{\beta}_g)\}^2}{np_k(\hat{\beta}_g)},$$

asymptotically follows a χ^2_{K-r-1} distribution with r indicating the dimensionality of β. More recently, Johnson (2004) took a Bayesian approach to constructing a χ^2 test statistic in the form of

$$Q^{\text{Bayes}}(\tilde{\beta}) = \sum_{k=1}^{K} \frac{\{m_k(\tilde{\beta}) - np_k\}^2}{np_k}, \tag{1.3}$$

where $\tilde{\beta}$ is a sample from the posterior distribution of β. In the Bayesian χ^2 test, the partition is constructed as follows. We prespecify $0 \equiv s_0 < s_1 < \cdots < s_K \equiv 1$, and let $p_k = s_k - s_{k-1}$, and let $m_k(\tilde{\beta})$ be the count of y_i's satisfying $F_{\tilde{\beta}}(y_i) \in [s_{k-1}, s_k)$, for $i = 1, \ldots, n$. Johnson (2004) showed that $Q^{\text{Bayes}}(\tilde{\beta})$ is asymptotically distributed as $\chi^2_{(K-1)}$ regardless of the dimensionality of β. Intuitively, by generating a posterior sample $\tilde{\beta}$, $Q^{\text{Bayes}}(\tilde{\beta})$ recovers the χ^2 distribution and the degrees of freedom that are lost due to computing the MLE of β. However, the Bayesian χ^2 test requires implementation of the usual Monte Carlo Markov chain (MCMC) procedure, which is computationally intensive and also depends on the prior distribution of β. In particular, the prior distribution on β must be noninformative. A major class of noninformative prior distributions are improper priors, which, however, may lead to improper posteriors. If some informative prior distribution is used for β, the asymptotic χ^2 distribution of $Q^{\text{Bayes}}(\tilde{\beta})$ may be distorted, i.e., $Q^{\text{Bayes}}(\tilde{\beta})$ is sensitive to the prior distribution of β. In addition, the Pearson-type statistic is largely based on the frequentist maximum likelihood approach, and thus combining a Bayesian posterior sample with the Pearson test is not natural. As a result, $Q^{\text{Bayes}}(\tilde{\beta})$ cannot be generally used in the classical maximum likelihood framework. Johnson (2007) further developed Bayesian model assessment using pivotal quantities along the similar direction in the Bayesian paradigm.

Our goal is to overcome the dependence of $Q^{\text{Bayes}}(\tilde{\beta})$ on the prior distribution and further expand the Pearson-type goodness-of-fit test to regression models in the classical maximum likelihood paradigm. We propose a bootstrap χ^2 test to evaluate model fitting, which is easy to implement, and does not require tedious computations other than calculating the MLE of the model parameter by fitting exactly the same model to a bootstrap sample of the original data. The new test statistic maintains the elegance of the Pearson-type formulation, as the right amount of randomness is produced as a whole set through a bootstrap sample to recover the classical χ^2 test. The proposed bootstrap χ^2 test does not require intensive MCMC sampling, and also it is more objective because it does not depend on any prior distribution. Moreover, it is more natural to combine the bootstrap procedure with the classical maximum likelihood estimation in the Pearson test, in contrast to using a posterior sample in the Bayesian paradigm.

The rest of this article is organized as follows. In Section 2, we propose the bootstrap χ^2 goodness-of-fit test using the MLE of the model parameter obtained from a bootstrap sample of the data, and derive the asymptotic distribution for the test statistic. In Section 3, we conduct simulation studies to examine the bootstrap χ^2 test in terms of the test size and statistical power, and also illustrate the proposed method using a real data example. Section 4 gives concluding remarks, and technical details are outlined in the appendix.

2. Pearson χ^2 test with bootstrap

Let (y_i, Z_i) denote the i.i.d. data for $i = 1, \ldots, n$, where y_i is the outcome of interest and Z_i is the r-dimensional covariate vector for subject i. For ease of exposition, we take the generalized linear model (GLM) to characterize the association between y_i and Z_i (McCullagh and Nelder, 1989). It is well known that GLMs are suitable for modeling a broad range of data structures, including both continuous and categorical data (e.g., binary or Poisson count data). We assume that the density function of y_i is from an exponential family in the form of

$$f(y_i|Z_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{2.1}$$

where θ_i is a location parameter, ϕ is a scalar dispersion parameter, and $a_i(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. The linear predictor $\eta_i = \beta^T Z_i$ can be linked with θ_i through a monotone differentiable function $h(\cdot)$, i.e., $\theta_i = h(\eta_i)$. This is a standard formulation of the GLM, with $E(y_i|Z_i) = b'(\theta_i)$ and $\text{Var}(y_i|Z_i) =$

$b''(\theta_i)a_i(\phi)$, where $b'(\cdot)$ and $b''(\cdot)$ represent the first and second derivatives, respectively.

We are interested in testing whether the model in (2.1) fit the observed data adequately. We illustrate the bootstrap χ^2 test under the GLM framework as follows. We first take a simple random sample with replacement from the observed data $\{(y_i, Z_i), i = 1, \ldots, n\}$, and denote the bootstrap sample as $\{(y_i^*, Z_i^*), i = 1, \ldots, n\}$. We then fit the original regression model to the bootstrap sample and obtain the MLE of β, denoted as β^*. We partition the range of $[0,1]$ into K intervals, $0 \equiv s_0 < s_1 < \cdots < s_K \equiv 1$, with $p_k = s_k - s_{k-1}$. Based on the original data $\{(y_i, Z_i), i = 1, \ldots, n\}$ and β^*, we then compute the Pearson-type bin counts for each partition. Let $m_k(\beta^*)$ denote the number of subjects satisfying $F_{\beta^*}(y_i|Z_i) \in [s_{k-1}, s_k)$, where $F_\beta(y_i|Z_i)$ is the cumulative distribution function corresponding to $f(y_i|Z_i)$ in (2.1). That is

$$m_k(\beta^*) = \sum_{i=1}^{n} I(s_{k-1} \le F_{\beta^*}(y_i|Z_i) < s_k)$$

and then we define

$$Q^{\text{Boot}}(\beta^*) = \sum_{k=1}^{K} \frac{\{m_k(\beta^*) - np_k\}^2}{np_k}. \tag{2.2}$$

The proposed bootstrap chi-squared statistic $Q^{\text{Boot}}(\beta^*)$ has the following asymptotic property.

Theorem 2.1. *Under the regularity conditions in the appendix, $Q^{\text{Boot}}(\beta^*)$ asymptotically converges to a chi-squared distribution with $K-1$ degrees of freedom, $\chi^2_{(K-1)}$, under the null hypothesis.*

We outline the key steps of the proof in the appendix. For continuous distributions, $m_k(\beta^*)$ in (2.2) can be obtained in a straightforward way. However, if the data are from a discrete distribution, the corresponding distribution function $F(\cdot)$ is a step function. In this case, we replace the step function with a piecewise linear function that connects the jump points, and redefine $F_{\beta^*}(y_i|Z_i)$ to be a uniform distribution between the two adjacent endpoints of the line segment. In particular, for binary data we define

$$\pi_i^* = \frac{1}{1 + \exp(\beta^{*T} Z_i)},$$

where β^* is the MLE for a bootstrap sample under the logistic regression. If $y_i = 0$, then we take $F_{\beta^*}(y_i|Z_i)$ to be a uniform draw from $(0, \pi_i^*)$; and if $y_i = 1$, we take $F_{\beta^*}(y_i|Z_i)$ to be a uniform draw from $(\pi_i^*, 1)$. In the Poisson regression, for each given subject with (y_i, Z_i), we can calculate the Poisson mean $\mu_i^* = \exp(\beta^{*T} Z_i)$ based on the bootstrap sample MLE β^*. We then take $F_{\beta^*}(y_i|Z_i)$ as a uniform draw from (π_i^L, π_i^U), where

$$\pi_i^L = \sum_{j=0}^{y_i - 1} \frac{\exp(-\mu_i^*)\mu_i^{*j}}{j!}, \quad \text{and} \quad \pi_i^U = \sum_{j=0}^{y_i} \frac{\exp(-\mu_i^*)\mu_i^{*j}}{j!}.$$

In the proposed goodness-of-fit test, the MLE of β needs to be calculated only once based on one bootstrap sample, and thus computation is not heavier than the classical Pearson chi-squared test. On the other hand, the test result depends on one particular bootstrap sample, which can be different for different bootstrap samples. Ideally, we may eliminate the randomness by calculating $E\{Q^{\text{Boot}}(\beta_b^*)|\text{data}\}$, where the expectation is taken over all the bootstrap samples conditional on the original data. In practice, we may take a large number of bootstrap samples, and for each of them we construct a chi-squared test statistic. Although these chi-squared values are correlated, the averaged chi-squared test statistic may provide an approximation to $E\{Q^{\text{Boot}}(\beta_b^*)|\text{data}\}$.

In terms of empirical distribution functions, Durbin (1973) and Stephens (1978) studied the half-sample method and random substitution for goodness-of-fit tests for distributional assumptions. In particular, using the randomly chosen half of the samples without replacement, the same distribution can be obtained as if the true parameters are known. Nevertheless, our bootstrap procedure not only examines the distributional assumptions, but it also checks the mean structure of the model.

3. Numerical studies

3.1. Simulations

We carried out simulation studies to examine the finite sample properties of the proposed bootstrap χ^2 goodness-of-fit test. We focused on the GLMs by simulating data from the linear model, the Poisson regression model, and the logistic model, respectively. We took the number of partitions $K = 5$ and the sample sizes $n = 50$, 100, and 200. For each model, we independently generated two covariates: the first covariate Z_1 was a continuous variable from the standard normal distribution and the second Z_2 was a Bernoulli variable taking a value of 0 or 1 with an equal probability of 0.5. We set the intercept $\beta_0 = 0.2$, and the two slopes corresponding to Z_1 and Z_2, $\beta_1 = 0.5$ and $\beta_2 = -0.5$. Under the linear regression model,

$$y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \epsilon, \tag{3.1}$$

we simulated the error term from a normal distribution with mean zero and variance 0.01 under the null hypothesis. The Poisson log-linear regression model took the form of

$$\log\mu = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2,$$

where Z_1 and Z_2 were generated in the same way as those in the linear model. The logistic model assumed the success probability p in the form of

$$\text{logit}(p) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2,$$

and all the rest of setups are the same as before. We conducted 1,000 simulations under each configuration.

Table 1

Test sizes of the proposed bootstrap χ^2 goodness-of-fit test with $K = 5$ at different significance levels of α, under the null hypothesis: linear, Poisson and logistic models, respectively

Model	n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.5$
Linear	50	0.013	0.047	0.095	0.249	0.522
	100	0.006	0.057	0.094	0.239	0.483
	200	0.006	0.038	0.096	0.245	0.483
Poisson	50	0.012	0.048	0.010	0.267	0.557
	100	0.010	0.052	0.109	0.250	0.479
	200	0.005	0.042	0.091	0.245	0.488
Logistic	50	0.014	0.047	0.098	0.286	0.542
	100	0.010	0.051	0.102	0.260	0.512
	200	0.009	0.058	0.104	0.253	0.495

The simulation results evaluating the test levels are summarized in Table 1. We can see that for each of the five prespecified significance levels of $\alpha = 0.01$ up to 0.5, the bootstrap χ^2 test clearly maintains the type I error rate under each model. As the sample size increases, the test sizes become closer to the corresponding nominal levels. Figure 1 exhibits the quantile-quantile (Q-Q) plots under each modeling structure with $n = 100$. Clearly, the proposed bootstrap χ^2 test recovers the χ^2 distribution, as all of the Q-Q plots using the MLE from a bootstrap sample closely match the straight diagonal lines. This demonstrates that the proposed bootstrap χ^2 test performed well with finite sample sizes. We also computed the classical Pearson test statistic when using the MLE calculated from the original data. The corresponding Q-Q plots are presented in Figure 1 as well. The Pearson test statistics using the original data MLE are lower than the expected $\chi^2_{(4)}$ quantiles. This confirms the findings by Chernoff and Lehmann (1954) and also extends their conclusions for the i.i.d. case to the general regression models.

We further examined the power of the proposed bootstrap χ^2 test by simulating data from the alternative hypothesis. Under the linear model, we simulated the error terms from a student $t_{(2)}$ distribution with two degrees of freedom, i.e., $\epsilon \sim t_{(2)}$ in model (3.1). The covariates were generated similarly to those in the null case. We took the number of partitions $K = 5$, the sample size $n = 150$, and conducted 1,000 simulations. Under the linear model with the $t_{(2)}$ error, the power of our χ^2 test was 0.893. In another simulation with the linear model, we generated data from an alternative model with an extra quadratic term of covariate Z_1, that is,

$$y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \gamma Z_1^2 + \epsilon,$$

while the null model is still given by (3.1). The power of the proposed χ^2 test was 0.817 for $\gamma = 0.15$, and 0.940 for $\gamma = 0.2$.

Similarly, for the Poisson regression model, we added an extra quadratic term in the Poisson mean function under the alternative model, that is,

$$\mu = \exp(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \gamma Z_1^2).$$

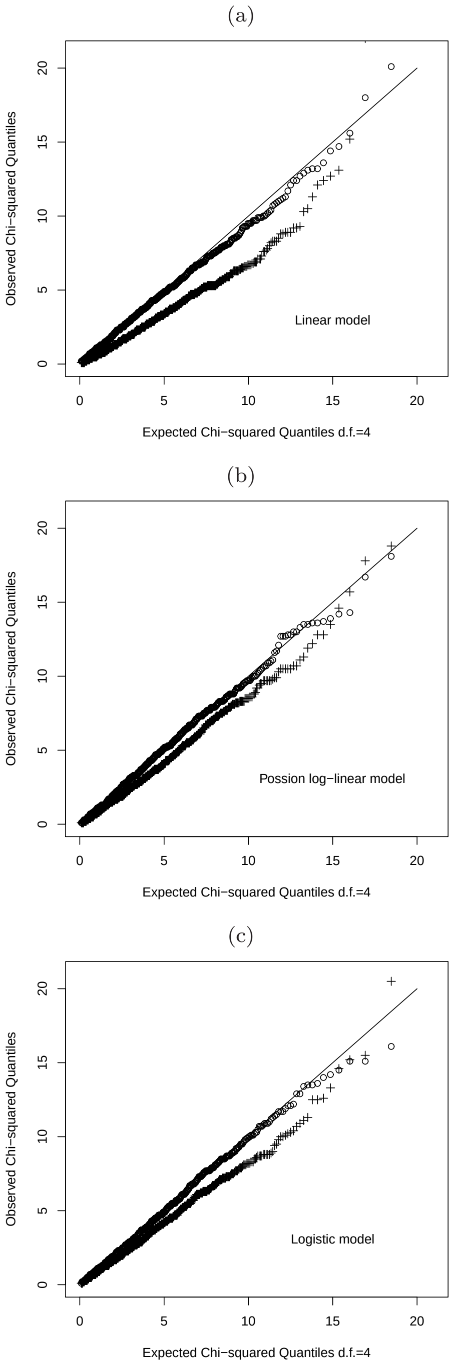


FIG 1. *Quantile-quantile plots for the bootstrap χ^2 test statistics with sample size $n = 100$ and $K = 5$ ("circle" representing the proposed $\chi^2_{(K-1)}$ statistics based on the bootstrap sample MLE, and "+" representing the classical Pearson statistics based on the original data MLE): (a) the linear regression model; (b) the Poisson log-linear model; and (c) the logistic model.*

If $\gamma = 0.5$, the power of our χ^2 test was 0.829, and if $\gamma = 0.6$, the power increased to 0.962. We also examined the case where the alternative model was from a negative binomial distribution but with the same mean as that of the Poisson mean. In particular, we took the mean of the negative binomial distribution $\mu = \exp(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$ and the negative binomial parameter $p = r/(r + \mu)$. The probability mass function of the negative binomial distribution is given by

$$P(x|p,r) = \binom{r + x - 1}{x} p^r (1-p)^x, \quad x = 0, 1, 2, \ldots,$$

which converges to a Poisson distribution (the null model), as $r \to \infty$. When $r = 0.7$, the power of our chi-squared test was 0.838, and when $r = 0.8$, the corresponding power was 0.783.

Finally, we examined the test power for the logistic regression model using the proposed χ^2 test. Under the alternative hypothesis, we added a quadratic term in the logistic model,

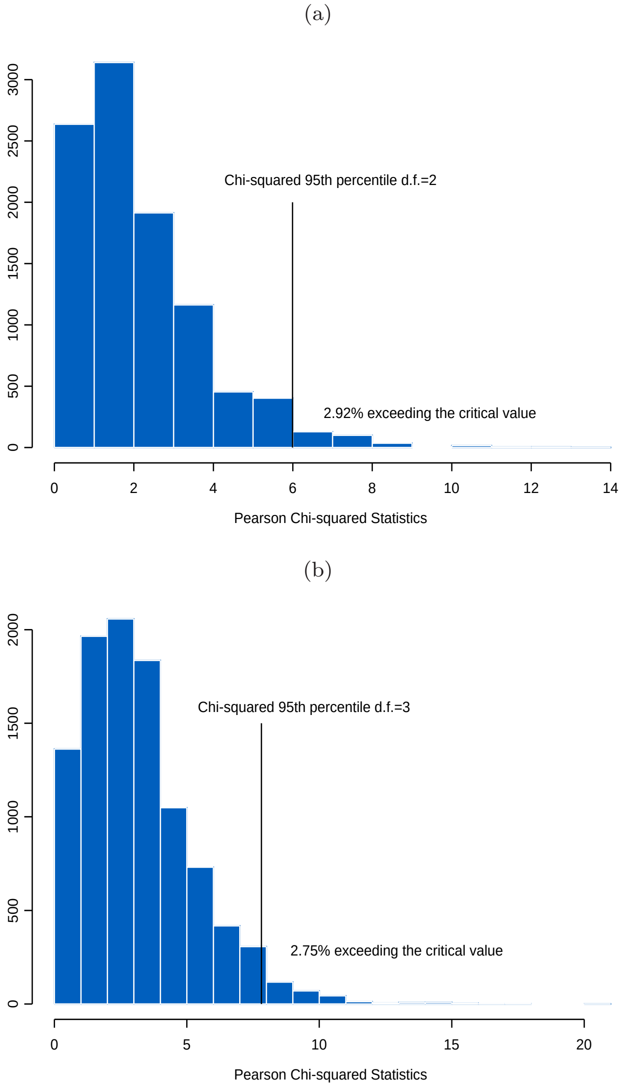
$$p = \frac{\exp(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \gamma Z_1^2)}{1 + \exp(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \gamma Z_1^2)},$$

where Z_1 was simulated from a uniform distribution on $(1, 2)$ and Z_2 was still a binary covariate. As $\gamma = 0$ corresponded to the null model, we took $\gamma = 0.4$ to yield a power of 0.897 for our test, and $\gamma = 0.5$ to have a power of 0.976. We also examined a different modeling structure by taking $p = \Psi(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$, where $\Psi(\cdot)$ is the cumulative distribution function of an exponential distribution. Under this alternative, the power of the proposed test was 0.929. Our test uses the bootstrap data MLE, which recovers the chi-squared distribution. In contrast, the Chernoff and Lehmann test statistic does not follow any explicit distribution.

3.2. Application

As an illustration, we applied the proposed goodness-of-fit test to a well-known steam data set described in Draper and Smith (1998). The steam study contained $n = 25$ observations measured at intervals from a steam plant. The outcome variable was the monthly use of steam, and the covariates of interest included the operating days per month and the average atmospheric temperature. The steam data set was analyzed using a linear regression model, which involved three unknown regression parameters and the variance of the errors. The linear model was claimed to be of adequate fit based on the plot of residuals versus the predicted outcomes. This was also confirmed by the Durbin-Watson test (Draper and Smith, 1998).

To quantify the model fit in a more objective way, we applied the proposed bootstrap χ^2 test to examine how well the linear model fit the data from the steam study. Because the sample size was quite small, we partitioned the range of $[0, 1]$ into 3 or 4 intervals, i.e., $K = 3$ or 4. We took 10,000 bootstrap samples from the original data, and for each of them, we computed the MLEs of


(a)

(b)

FIG 2. *Histograms of the Pearson-type goodness-of-fit test statistics for the steam data with (a) $K = 3$ and (b) $K = 4$.*

the model parameters. Based on these MLEs, we constructed our $\chi^2_{(K-1)}$ test statistics by plugging the bootstrap sample MLEs in the Pearson-type statistic. In Figure 2, we show the histograms of the proposed $\chi^2_{(2)}$ and $\chi^2_{(3)}$ statistics for $K = 3$ and 4, respectively. We can see that among 10,000 bootstrap χ^2 test statistics only 2.92% of the test statistics exceed the critical value at the significance level of $\alpha = 0.05$ for $K = 3$, while 2.75% for $K = 4$. Our findings provided strong evidence for the model fit, and thus confirmed that the linear regression model adequately fit the steam data.

4. Discussion

We have proposed a bootstrap-based modification to the classical Pearson χ^2 goodness-of-fit test for regression models, which is a major extension of the work of Chernoff and Lehmann (1954) and Johnson (2004). The new procedure replaces the classical MLE from the original data by the MLE from a bootstrap sample. Using the MLE of a bootstrap sample adjusts the right amount of randomness to the test statistic. Not only does the proposed method restore the degrees of freedom, but also the χ^2 distribution itself, which would have been a nonstandard distribution lying between two χ^2 distributions with different degrees of freedom. Our simulation studies have shown that the proposed test statistic performs well with small sample sizes, and increasingly so as the sample size increases.

Compared with the well-known Akaike information criterion (AIC) and Bayesian information criterion (BIC), we may use the averaged value of the chi-squared statistics computed from a large number of bootstrap samples for model selection or comparison. A smaller value of the averaged chi-squared statistic indicates a better fitting model. It is worth noting that there is no scale associated with the AIC and BIC statistics, thus they are not meaningful alone. In other words, the AIC and the BIC by themselves do not provide any information on the goodness-of-fit of a single model, and they are only interpretable when comparing two or more competing models. In contrast, not only can our averaged bootstrap χ^2 statistic be used for model comparison or model selection, but also it is closely related to the χ^2 distribution, and as an approximation, one would know how well a model fits the data based on the corresponding χ^2 distribution. That is, the proposed test can be used for both model diagnostic and model selection at the same time. For example, a very large value of the averaged χ^2_{K-1} value for a small K may shed doubt on the model fit.

For the i.i.d. data, the minimum χ^2 statistic estimates the unknown parameter β by minimizing the χ^2 statistic or maximizing the grouped-data likelihood (Cramér, 1946). The minimum χ^2 statistic may not be directly applicable in regression settings due to difficulties involved in grouping the data with regression models. Also, it is challenging to generalize the proposed bootstrap Pearson-type statistic to censored data with commonly used semiparametric Cox proportional hazards model in survival analysis (Cox, 1972; Akritas, 1988; and Akritas and Torbeyns, 1997). Future research is warranted along these directions.

Appendix A: Proof of Theorem 2.1

We assume the conditions (a)-(d) in Cramér (1946, pp. 426-427), and the regularity conditions in Chernoff and Lehmann (1954, p. 581). The conditions in Cramér (1946) are sufficient to prove the χ^2 distribution when using the grouped data MLE. We essentially require the likelihood to be a smooth function of the parameter, the information in the sample increases with the sample size, and the third-order (partial) derivatives of the density function exist. Let β_0 be the true value of the parameter β, let $\hat{\beta}$ be the MLE of β based on the original observations, and let β^* be the MLE of β based on the bootstrap sample. Denote

$$\begin{aligned} \hat{m}_k &= \sum_{i=1}^{n} I\{s_{k-1} \le F_{\hat{\beta}}(y_i|Z_i) < s_k\}, \\ m_k^* &= \sum_{i=1}^{n} I\{s_{k-1} \le F_{\beta^*}(y_i|Z_i) < s_k\}, \\ m_k &= \sum_{i=1}^{n} I\{s_{k-1} \le F_{\beta_0}(y_i|Z_i) < s_k\}. \end{aligned}$$

Let $G(\alpha, \gamma, s) = E[F_\gamma\{F_\alpha^{-1}(s|Z_i)|Z_i\}]$, define $\hat{s}_k = G(\beta_0, \hat{\beta}, s_k)$, $t_k = G(\hat{\beta}, \beta_0, s_k)$, $r_k = G(\beta^*, \beta_0, s_k)$, $\hat{p}_k = \hat{s}_k - \hat{s}_{k-1}$, and $b_k = t_k - t_{k-1}$.

We have that

$$\frac{m_k^* - np_k}{\sqrt{np_k}} = \frac{m_k^* - \hat{m}_k}{\sqrt{np_k}} + \frac{\hat{m}_k - m_k}{\sqrt{np_k}} + \frac{m_k - np_k}{\sqrt{np_k}}. \tag{A.1}$$

If we follow the notation of (5) in Chernoff and Lehmann (1954), then $(m_k - np_k)/\sqrt{np_k} = \epsilon_k$. We first analyze the term $(\hat{m}_k - m_k)/\sqrt{np_k}$, by writing

$$\begin{aligned} &\frac{\hat{m}_k - m_k}{\sqrt{n}} \\ =\ & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[I\{s_{k-1} \le F_{\hat{\beta}}(y_i|Z_i) < s_k\} - I\{s_{k-1} \le F_{\beta_0}(y_i|Z_i) < s_k\} \Big] \\ =\ & \sqrt{n} \Big[EI\{s_{k-1} \le F_{\hat{\beta}}(y_i|Z_i) < s_k\} - EI\{s_{k-1} \le F_{\beta_0}(y_i|Z_i) < s_k\} \Big] \\ & + O_p(n^{-1/2}). \end{aligned}$$

The remaining term is of the same order as the standard deviation of $I\{s_{k-1} \le F_{\hat{\beta}}(y_i|Z_i) < s_k\} - I\{s_{k-1} \le F_{\beta_0}(y_i|Z_i) < s_k\}$, which takes the value of 0 with probability $1 - O(\hat{\beta} - \beta_0)$, and the value of 1 or -1 with probability $O(\hat{\beta} - \beta_0) = O_p(n^{-1/2})$. Thus, we can further write

$$\begin{aligned} &\frac{\hat{m}_k - m_k}{\sqrt{n}} \\ =\ & \sqrt{n} \Big[\Pr\{F_{\hat{\beta}}(y_i|Z_i) < s_k\} - \Pr\{F_{\hat{\beta}}(y_i|Z_i) < s_{k-1}\} \\ & - \Pr\{F_{\beta_0}(y_i|Z_i) < s_k\} + \Pr\{F_{\beta_0}(y_i|Z_i) < s_{k-1}\} \Big] + O_p(n^{-1/2}) \\ =\ & \sqrt{n} \Big(\Pr\Big[F_{\beta_0}(y_i|Z_i) < F_{\beta_0}\{F_{\hat{\beta}}^{-1}(s_k|Z_i)|Z_i\} \Big] \\ & - \Pr\Big[F_{\beta_0}(y_i|Z_i) < F_{\beta_0}\{F_{\hat{\beta}}^{-1}(s_{k-1}|Z_i)|Z_i\} \Big] - s_k + s_{k-1} \Big) + O_p(n^{-1/2}) \\ =\ & \sqrt{n}(t_k - t_{k-1} - s_k + s_{k-1}) + O_p(n^{-1/2}) \\ =\ & \sqrt{n}(b_k - p_k) + O_p(n^{-1/2}). \end{aligned}$$

We now show that $b_k - p_k$ can be approximated by $p_k - \hat{p}_k$, in the classical MLE construction. Note that $s_k = G(\beta_0, \beta_0, s_k) = G(\hat{\beta}, \hat{\beta}, s_k)$. Denoting

$$G_1(\alpha, \gamma, s) = \frac{\partial G(\alpha, \gamma, s)}{\partial \alpha}, \quad G_2(\alpha, \gamma, s) = \frac{\partial^2 G(\alpha, \gamma, s)}{\partial \alpha \partial \gamma^T},$$

we have that

$$\begin{aligned}
& (b_k - p_k) - (p_k - \hat{p}_k) \\
= \;& G(\hat{\beta}, \beta_0, s_k) - G(\hat{\beta}, \beta_0, s_{k-1}) - G(\beta_0, \beta_0, s_k) + G(\beta_0, \beta_0, s_{k-1}) \\
& - G(\hat{\beta}, \hat{\beta}, s_k) + G(\hat{\beta}, \hat{\beta}, s_{k-1}) + G(\beta_0, \hat{\beta}, s_k) - G(\beta_0, \hat{\beta}, s_{k-1}) \\
= \;& G_1(\beta_0, \beta_0, s_k)^T(\hat{\beta} - \beta_0) - G_1(\beta_0, \beta_0, s_{k-1})^T(\hat{\beta} - \beta_0) \\
& - G_1(\beta_0, \hat{\beta}, s_k)^T(\hat{\beta} - \beta_0) + G_1(\beta_0, \hat{\beta}, s_{k-1})^T(\hat{\beta} - \beta_0) + O_p(n^{-1}) \\
= \;& -(\hat{\beta} - \beta_0)^T G_2(\beta_0, \beta_0, s_k)(\hat{\beta} - \beta_0) \\
& + (\hat{\beta} - \beta_0)^T G_2(\beta_0, \beta_0, s_{k-1})(\hat{\beta} - \beta_0) + O_p(n^{-1}) \\
= \;& O_p(n^{-1}).
\end{aligned}$$

Thus, $\sqrt{n}(b_k - p_k) = \sqrt{n}(p_k - \hat{p}_k) + O_p(n^{-1/2})$, and

$$\frac{\hat{m}_k - m_k}{\sqrt{np_k}} = \frac{n(p_k - \hat{p}_k)}{\sqrt{np_k}} + o_p(1) = -\hat{v}_k + o_p(1),$$

where $\hat{v}_k$ is defined in (7) of Chernoff and Lehmann (1954).

We now consider the first term in (A.1). Following the bootstrap principle, the conditional distribution of this term should be the same as that of the second one. We show that in fact the two terms are identically distributed as $n \to \infty$, and they are independent. As an intermediate result, we have already established that

$$\frac{\hat{m}_k - m_k}{\sqrt{n}} = \{G_1(\beta_0, \beta_0, s_k) - G_1(\beta_0, \beta_0, s_{k-1})\}^T \sqrt{n}(\hat{\beta} - \beta_0) + o_p(1).$$

Following a similar derivation, we have that

$$\begin{aligned}
& \frac{m_k^* - \hat{m}_k}{\sqrt{n}} \\
= \;& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[I\{s_{k-1} \le F_{\beta^*}(y_i|Z_i) < s_k\} - I\{s_{k-1} \le F_{\hat{\beta}}(y_i|Z_i) < s_k\} \Big] \\
= \;& \sqrt{n} \Big[EI\{s_{k-1} \le F_{\beta^*}(y_i|Z_i) < s_k\} - EI\{s_{k-1} \le F_{\hat{\beta}}(y_i|Z_i) < s_k\} \Big] \\
& + O_p(n^{-1/2}) \\
= \;& \sqrt{n} \Big[\Pr\{F_{\beta^*}(y_i|Z_i) < s_k\} - \Pr\{F_{\beta^*}(y_i|Z_i) < s_{k-1}\} \\
& - \Pr\{F_{\hat{\beta}}(y_i|Z_i) < s_k\} + \Pr\{F_{\hat{\beta}}(y_i|Z_i) < s_{k-1}\} \Big] + O_p(n^{-1/2})
\end{aligned}$$

$$\begin{aligned}
&= \sqrt{n}\Big[\Pr\{F_{\beta_0}(y_i|Z_i) < F_{\beta_0}\{F^{-1}_{\beta^*}(s_k|Z_i)|Z_i\} \\
&\quad - \Pr\{F_{\beta_0}(y_i|Z_i) < F_{\beta_0}\{F^{-1}_{\beta^*}(s_{k-1}|Z_i)|Z_i\} \\
&\quad - \Pr\{F_{\beta_0}(y_i|Z_i) < F_{\beta_0}\{F^{-1}_{\hat\beta}(s_k|Z_i)|Z_i\} \\
&\quad + \Pr\{F_{\beta_0}(y_i|Z_i) < F_{\beta_0}\{F^{-1}_{\hat\beta}(s_{k-1}|Z_i)|Z_i\}\Big] + O_p(n^{-1/2}) \\
&= \sqrt{n}(r_k - r_{k-1} - t_k + t_{k-1}) + O_p(n^{-1/2}) \\
&= \sqrt{n}\{G(\beta^*, \beta_0, s_k) - G(\beta^*, \beta_0, s_{k-1}) - G(\hat\beta, \beta_0, s_k) + G(\hat\beta, \beta_0, s_{k-1})\} \\
&\quad + O_p(n^{-1/2}) \\
&= \sqrt{n}\{G_1(\hat\beta, \beta_0, s_k)(\beta^* - \hat\beta) - G_1(\hat\beta, \beta_0, s_{k-1})(\beta^* - \hat\beta)\} + O_p(n^{-1/2}) \\
&= \{G_1(\beta_0, \beta_0, s_k) - G_1(\beta_0, \beta_0, s_{k-1})\}\sqrt{n}(\beta^* - \hat\beta) + o_p(1).
\end{aligned}$$

Note that $G_1(\beta_0, \beta_0, s_k) - G_1(\beta_0, \beta_0, s_{k-1})$ is a nonrandom quantity. As $n \to \infty$, $\sqrt{n}(\hat\beta - \beta_0)$ converges to a normal distribution with mean zero. Conditional on $\hat\beta$, $\sqrt{n}(\beta^* - \hat\beta)$ also converges to a mean zero normal distribution. In addition, $\sqrt{n}(\beta^* - \hat\beta)$ and $\sqrt{n}(\hat\beta - \beta_0)$ are asymptotically uncorrelated, so they are independent of each other asymptotically. Hence, we can represent the first term of (A.1) as $v_k^* + o_p(1)$, which is independent of $\hat v_k$ and ϵ_k, and has the same distribution as $\hat v_k$.

Let $\epsilon = (\epsilon_1, \dots, \epsilon_K)^T$, and similarly define $\hat v$ and v^*. Now, following the notations and arguments of Chernoff and Lehmann (1954), we let the information matrix be $\tilde J = D^T D$, where D is the matrix with element $(\partial p_k/\partial \beta_j)/\sqrt{p_k}$ for $j, k = 1, \dots, K$. Note that $\epsilon \sim N(0, I - qq^T)$ asymptotically, where $q = (\sqrt{p_1}, \dots, \sqrt{p_K})^T$,

$$\hat v = D(\tilde J + J^*)^{-1}D^T\epsilon + D(\tilde J + J^*)^{-1}\eta + o_p(1),$$

where $\eta \sim N(0, J^*)$, J^* is defined the same as in Chernoff and Lehmann (1954, p. 583), and η is independent of ϵ. We use e and τ to denote random variables that have the same distributions as ϵ and η, respectively. Note that ϵ, η, e and τ are all independent of each other. We then have that

$$\begin{aligned}
&\left(\frac{m_1^* - np_1}{\sqrt{np_1}}, \dots, \frac{m_K^* - np_K}{\sqrt{np_K}}\right)^T \\
&= \epsilon - \hat v - v^* + o_p(1) \\
&= \epsilon - D(\tilde J + J^*)^{-1}D^T\epsilon - D(\tilde J + J^*)^{-1}\eta - D(\tilde J + J^*)^{-1}D^T e \\
&\quad - D(\tilde J + J^*)^{-1}\tau + o_p(1) \\
&= \{I - D(\tilde J + J^*)^{-1}D^T\}\epsilon - D(\tilde J + J^*)^{-1}\eta - D(\tilde J + J^*)^{-1}D^T e \\
&\quad - D(\tilde J + J^*)^{-1}\tau + o_p(1). \qquad \text{(A.2)}
\end{aligned}$$

Note that $D^T q = 0$, $\text{var}(\eta) = J^*$ and $D^T D = \tilde J$. As $n \to \infty$, (A.2) converges to a normal random vector with the variance-covariance matrix

$$\begin{aligned}
&\{I - D(\tilde J + J^*)^{-1}D^T\}(I - qq^T)\{I - D(\tilde J + J^*)^{-1}D^T\} \\
&+ D(\tilde J + J^*)^{-1}J^*(\tilde J + J^*)^{-1}D^T
\end{aligned}$$

$$
\begin{aligned}
& + D(\tilde{J}+J^*)^{-1}D^T(I-qq^T)D(\tilde{J}+J^*)^{-1}D^T \\
& + D(\tilde{J}+J^*)^{-1}J^*(\tilde{J}+J^*)^{-1}D^T \\
= \;& I - qq^T - 2D(\tilde{J}+J^*)^{-1}D^T + 2D(\tilde{J}+J^*)^{-1}J^*(\tilde{J}+J^*)^{-1}D^T \\
& + 2D(\tilde{J}+J^*)^{-1}\tilde{J}(\tilde{J}+J^*)^{-1}D^T \\
= \;& I - qq^T,
\end{aligned}
$$

which is the same as the asymptotic variance-covariance matrix of ϵ. This completes the proof that $Q^{\mathrm{Boot}}(\beta^*)$ has a $\chi^2_{(K-1)}$ distribution as $n \to \infty$.

Acknowledgements

We thank Professor Valen Johnson for many helpful and stimulating discussions, and also thank referees for their comments. Yin's research was supported by a grant (Grant No. 784010) from the Research Grants Council of Hong Kong. Ma's research was supported by a US NSF grant.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds.), pp. 267–281. Budapest: Akademiai Kiado. MR0483125

Akritas, M. G. (1988). Pearson-type goodness-of-fit tests: the univariate case. *Journal of the American Statistical Association* **83**, 222–230. MR0941019

Akritas, M. G. and Torbeyns, A. F. (1997). Pearson-type goodness-of-fit tests for regression. *The Canadian Journal of Statistics* **25**, 359–374. MR1486917

Ali, M. M. (1984). An approximation to the null distribution and power of the Durbin-Watson statistic. *Biometrika* **71**, 253–261. MR0767153

Chernoff, H. and Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics* **25**, 579–586. MR0065109

Claeskens, G. and Hjort, L. N. (2008). *Model Selection and Model Averaging*. Cambridge University Press. MR2431297

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220. MR0341758

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press. MR0016588

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. 3rd edition, John Wiley & Sons: New York. MR1614335

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function.* SIAM Publications No. 9. Philadelphia. MR0305507

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**, 419–426. MR0137199

Johnson, V. E. (2004). A Bayesian χ^2 test for goodness-of-fit. *The Annals of Statistics* **32**, 2361–2384. MR2153988

Johnson, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Analysis* **2**, 719–734. MR2361972

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall. MR0727836

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* **50**, 157–175.

Stephens, M. A. (1978). On the half-sample method for goodness-of-fit. *Journal of the Royal Statistical Society, Series B* **40**, 64–70.

Stute, W. and Zhu, L.-X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics* **29**, 535–545. MR1925573

Stute, W., Manteiga, G. W. and Quindimilm, P. M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association* **93**, 141–149. MR1614600

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association* **86**, 420–426. MR1137124