A New Bootstrap Goodness-of-Fit Test for Normal Linear Regression Models

Scott H. Koeneman[a,*], Joseph E. Cavanaugh[b]

[a] *Division of Biostatistics and Bioinformatics, Thomas Jefferson University, 130 S. 9th St., Philadelphia, PA, 55038*
[b] *Department of Biostatistics, University of Iowa, 145 N. Riverside Dr., Iowa City, IA, 52242*

Abstract

In this work, the distributional properties of the goodness-of-fit term in likelihood-based information criteria are explored. These properties are then leveraged to construct a novel goodness-of-fit test for normal linear regression models that relies on a non-parametric bootstrap. Several simulation studies are performed to investigate the properties and efficacy of the developed procedure, with these studies demonstrating that the bootstrap test offers distinct advantages as compared to other methods of assessing the goodness-of-fit of a normal linear regression model. Our inferential technique can be employed using the `DBModelSelect` R package, available freely via the Comprehensive R Archive Network.

Keywords: Information Criteria, Information Matrix Test, Normal Distribution, Resampling, Robust Variance

*Corresponding author
Email address: `Scott.Koeneman@jefferson.edu` (Scott H. Koeneman)

1. Introduction

1.1. Goodness-of-Fit

Broadly, the term goodness-of-fit as it pertains to statistical modeling refers to the degree to which a certain model and its associated assumptions align with the observed data. The notion of goodness-of-fit is relevant in the model selection realm of information criteria, particularly to the Akaike information criterion [1]. Using $\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ to denote the log-likelihood of the fitted model and p to denote the number of functionally independent estimated parameters, the Akaike information criterion (AIC) for a fitted model can be expressed as

$$AIC = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + 2p.$$

The $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ term based on the empirical log-likelihood is known as the goodness-of-fit term as it represents the degree to which the fitted model conforms to the observed data $\boldsymbol{y}$ [3]. This term is also present in other likelihood-based information criteria such as the Bayesian information criterion [15]. The goodness-of-fit term will only grow smaller, indicating better conformity to the data, as complexity is added to the model. Thus, this goodness-of-fit term ironically does not encompass full goodness-of-fit considerations, and other considerations are needed to avoid overfitting and to detect violations of model assumptions.

Linear regression analysis imposes a number of assumptions about the data at hand, among them conditional independence and linearity, independence of errors that are normally distributed with a mean of zero, and homoskedasticity, that being constant error variance. While this work will focus on assessments of the goodness-of-fit and assumptions of traditional normal linear regression models, it is worth noting that assessments of goodness-of-fit exist for many other modeling paradigms as well, and many of the same principles apply.

Residual plots serve as a subjective visual method for assessing the appropriateness of a fitted linear regression model [12]. Across different values of the predictors and observed values, the residuals should not exhibit any specific pattern other than a mean of zero and constant variance if the assumptions of the

linear regression model are met. Thus, deviations from this pattern in the form of a residual plot that exhibits curvature, or a change in spread, can indicate violations to linearity and homoskedasticity, respectively.

However, this method of visual inspection carries with it certain limitations. When there are a large number of predictors present in a model, it may be impractical to visually inspect every possible residual plot with any degree of scrutiny. If heteroskedasticity is induced by covariates that have not been observed, a visual inspection will likely not reveal this violation of assumptions.

Hypothesis tests for the goodness-of-fit of a linear model offer an alternative to visual methods. These tests posit a null hypothesis that the model adequately accommodates the data and that the associated assumptions are satisfied, with an alternative that the assumptions are violated in some fundamental way. One such test is the Breusch-Pagan test, which postulates a null hypothesis that a linear regression model does not violate homoskedasticity [2]. This test involves performing an auxiliary linear regression on a transformation of the squared residuals from the candidate linear regression model against the covariates of interest.

However, one limitation of the Breusch-Pagan test is that it is only designed to detect a relationship between the covariates and the squared residuals that is linear, and thus it will not produce efficacious results if the heteroskedasticity present is not linear [18]. An alternative to the Breusch-Pagan test in this regard is the White test for homoskedasticity, which shares the same null hypothesis of homoskedasticity of a linear model as the Breusch-Pagan test [16]. The White test produces a test statistic that is sensitive to deviations to the null hypothesis in the form of heteroskedasticity related to the squares and cross products of regressors. Additionally, the test may indicate misspecification of the model if the cross products of certain regressors should be included in the model, but are not.

Another type of goodness-of-fit test is found in the information matrix test. The information matrix test compares the differences between the elements of the observed information matrix and the elements of the outer product of

the score vector, with a null hypothesis that the model is correctly specified [17]. This test is more general than the tests for heteroskedasticity, as the null hypothesis can be violated in a number of fashions. While able to detect misspecification of various kinds, the information matrix test can struggle with power in certain scenarios.

As the original formulation of the information matrix test can be computationally difficult to implement, an alternate method exists using an auxiliary regression to assist in calculation of the test statistic [4]. However, both the classical and regression variants of the information matrix test can struggle mightily with their performance in small sample sizes. Thus, another method of performing the test involves employing a parametric bootstrap to aid in calculation of the test statistic [5]. This method can be shown to better maintain a desired test level in small to moderate sample sizes.

The various ways in which goodness-of-fit tests can detect misspecification bring to attention a weakness of the hypothesis test methods as opposed to methods of visual inspection. When a hypothesis is rejected, we can be reasonably confident that an assumption is violated; however, by simply performing the test, we do not glean much information as to how exactly the model may be misspecified. This is in contrast to the method of visually observing residual plots wherein specific issues may be easier to identify.

In addition, both the White and Breusch-Pagan tests cannot detect heteroskedasticity induced by unobserved covariates, as their auxiliary regressions only use what has been observed. Such heteroskedasticity may not affect the bias and consistency of effect estimates, but may lead to a loss of efficiency of estimates.

1.2. Robust Variance Estimation

When employing likelihood theory to obtain parameter estimates for a model, one can often leverage the properties of maximum likelihood estimators (MLEs) to produce reasonable estimates of the variance of the estimators, and thus perform inference [13]. However, these properties are only guaranteed to hold when

the model is properly specified, and inefficient or biased estimates may result when assumptions do not hold. This has given rise to robust variance estimators that rely on fewer assumptions than classical likelihood theory, yet can still quantify the variability of a statistic at hand.

Huber [8] first provided justifications for consistency and asymptotic normality of maximum likelihood estimators under conditions weaker than had been previously shown. White [17] expanded upon this notion by deriving covariance matrix estimates for maximum likelihood estimators that are robust to a variety of different types of misspecification and only assume conditional independence and certain other regularity conditions. Defining $\boldsymbol{\theta}_*$ as the pseudo-true parameter and $\hat{\boldsymbol{\theta}}_n$ as the MLE for a sample of size n, White established the asymptotic relation

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \xrightarrow{d} N(0, \boldsymbol{C}(\boldsymbol{\theta}_*)).$$

The large sample covariance matrix $\boldsymbol{C}(\boldsymbol{\theta})$ can be defined using two positive definite matrices $\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{\theta})$, where the (i, j) element of $\boldsymbol{A}(\boldsymbol{\theta})$ is defined as

$$\boldsymbol{A}(\boldsymbol{\theta})_{i,j} = E\left[\frac{\partial^2 f(\boldsymbol{y}, \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]$$

and the (i, j) element of $\boldsymbol{B}(\boldsymbol{\theta})$ is defined as

$$\boldsymbol{B}(\boldsymbol{\theta})_{i,j} = E\left[\frac{\partial f(\boldsymbol{y}, \boldsymbol{\theta})}{\partial\theta_i}\frac{\partial f(\boldsymbol{y}, \boldsymbol{\theta})}{\partial\theta_j}\right]$$

where $i = 1, ..., p$ and $j = 1, ..., p$ respectively. These matrices then combine to define $\boldsymbol{C}(\boldsymbol{\theta})$ as

$$\boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})^{-1},$$

and thus evaluating this quantity at $\boldsymbol{\theta}_*$, one arrives at the robust asymptotic variance estimate. The resulting estimator, and those that are similar in form, are often called sandwich variance estimators due to one quantity being sandwiched in between two identical others to form the statistic.

As in general one will not know the pseudo-true parameter $\boldsymbol{\theta}_*$, White formulates the matrices $\boldsymbol{A_n}(\boldsymbol{\theta})$ and $\boldsymbol{B_n}(\boldsymbol{\theta})$, with the (i, j) element of $\boldsymbol{A_n}(\boldsymbol{\theta})$ defined as

$$\boldsymbol{A_n}(\boldsymbol{\theta})_{i,j} = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 f(y_t, \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}$$

and the (i, j) element of $\boldsymbol{B_n}(\boldsymbol{\theta})$ as

$$\boldsymbol{B_n}(\boldsymbol{\theta})_{i,j} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial f(y_t, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial f(y_t, \boldsymbol{\theta})}{\partial \theta_j}.$$

White proposes calculating $\boldsymbol{A_n}(\boldsymbol{\theta})$ and $\boldsymbol{B_n}(\boldsymbol{\theta})$ from the data, and evaluating them at the MLE, to produce

$$\boldsymbol{C_n}(\hat{\boldsymbol{\theta}}) = \boldsymbol{A_n}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{B_n}(\hat{\boldsymbol{\theta}})\boldsymbol{A_n}(\hat{\boldsymbol{\theta}})^{-1}.$$

It can then be shown that

$$\boldsymbol{C_n}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{a.s.} \boldsymbol{C}(\boldsymbol{\theta}_*).$$

Thus, we have a variance estimator for the MLE that is both robust to model misspecification and can be calculated using the data, but yet also will be approximately equivalent to the standard likelihood theory estimator if the model is correctly specified. This sandwich estimator, and others like it, can be used to perform inference related to the parameters if one calls into question the strong assumptions involved in using the traditional maximum likelihood estimator. However, if these assumptions cannot be met and the model appears to be misspecified, one may ponder the merit of performing inference on the pseudo-true parameters in the first place. Therefore, robust variance estimators serve as a hedge against slight deviances from a model being correctly specified, not as a tool that remediates poor model selection.

2. Derivations and Test Formulation

We will first explore the variance of the log-likelihood goodness-of-fit term present in likelihood-based information criteria. We will assume a scenario where a normal linear regression model is being fit to the data of interest, and that this model is not misspecified. Thus, this model is of the proper parametric family and contains the requisite mean structure, although the mean structure may contain extraneous variables in the case of an overspecified model.

Assuming a linear model has been fit using maximum likelihood with fitted parameters $\hat{\boldsymbol{\theta}}$, the goodness-of-fit term can be decomposed as

$$-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = n\log(2\pi) + n + n\log(\hat{\sigma}^2), \tag{1}$$

where $\hat{\sigma}^2$ denotes the maximum likelihood estimate for the error variance σ^2. Note that the only term here that is random is the statistic $n\log(\hat{\sigma}^2)$. Thus, if we can quantify the variability of this term, we can quantify the variability of the entire goodness-of-fit term.

To achieve this end, we first consider $\hat{\sigma}^2$. As the linear model is assumed to not be misspecified, and $\hat{\sigma}^2$ is a maximum likelihood estimator, this estimator will have an asymptotic variance related to the inverse of the Fisher information [6]. The Fisher information as it relates to the parameter vector $\boldsymbol{\theta'} = [\boldsymbol{\beta'}, \sigma^2]$, where $\boldsymbol{\beta}$ represents the regression coefficients present in the model, can be shown to be

$$-E\left[\frac{\partial^2 \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})}{\partial \boldsymbol{\theta}^2}\right] = \boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\boldsymbol{X'X}}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix},$$

where $\boldsymbol{X}$ is the design matrix of the regression, and 0 is a vector of zeroes. Thus, we may take the inverse of the Fisher information matrix and isolate the element related to the error variance σ^2. We see that this element will be

$$\boldsymbol{\mathcal{I}}^{-1}(\sigma^2) = \frac{2\sigma^4}{n}.$$

Using the above relation, and applying the property of asymptotic normality of the MLE $\hat{\sigma}^2$ in this case of a properly specified normal linear regression model, we see that the asymptotic distribution

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$$

should hold.

To find the variance of $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$, we must find the variance of $n\log(\hat{\sigma}^2)$. Additionally, the above asymptotic distribution involves the true σ^2 to which we will not have access in practical scenarios.

We will address both of these issues by employing the delta method [14]. We propose a transformation of the form

$$g(x) = \log(x).$$

Thus, applying the delta method to our above asymptotic distribution with $g(x)$ as the function of interest, we see that

$$\sqrt{n}(\log(\hat{\sigma}^2) - \log(\sigma^2)) \xrightarrow{d} N(0, 2).$$

Armed with the above asymptotic relationship and assuming that this asymptotic property approximately holds in a setting with a finite n, we have that

$$n \log(\hat{\sigma}^2) \dot{\sim} N\left(n \log(\sigma^2), 2n\right).$$

Thus, assuming that the model is appropriately specified, the variance of $n \log(\hat{\sigma}^2)$ will be approximately $2n$. Applying this variance back to the goodness-of-fit term, we see that the approximation

$$Var\left[-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right] \approx 2n$$

is justified. Furthermore, $2n$ could also serve as an approximation to the variance of AIC or BIC for this correctly specified linear regression model. This approximation has the same form no matter the complexity of the design matrix $\boldsymbol{X}$ or value of the true parameters $\boldsymbol{\theta}$, making it useful as a general tool.

It should be noted that this approximation does rely on asymptotic properties. In the Appendix, an exact variance for $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ is found which does not rely on asymptotic properties. However, this variance is more complicated to compute than the simple approximation $2n$, and was not found to provide any meaningful benefit over $2n$ when used in procedures developed later in this work.

With the previous derivation in hand, we now develop an estimator for the variance of $n \log(\hat{\sigma}^2)$, and thus $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$, that need not assume a given fitted normal linear regression model is correctly specified.

Assume that we once again fit a linear regression model with parameters $\boldsymbol{\theta}' = [\boldsymbol{\beta}', \sigma^2]$ to our data, and suppose we do not know whether this model is correctly specified. We wish to construct an estimator for $Var\left[-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]$ that is robust to model misspecification. This estimator will be constructed using the White robust sandwich variance estimator, employing much of the notation related to this development that was introduced in the first section of this work [16].

We let $\boldsymbol{I_n}(\boldsymbol{\theta})$ denote the observed information matrix with regards to our specified linear regression model. With $\boldsymbol{X}$ denoting the n by r design matrix, we see that the quantity $\boldsymbol{A_n}(\boldsymbol{\theta})$ used in the White estimator is found to be

$$\boldsymbol{A_n}(\boldsymbol{\theta}) = \frac{1}{n}\begin{bmatrix} \frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2} & \left[\frac{-(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{X}}{\sigma^4}\right]' \\ \left[\frac{-(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{X}}{\sigma^4}\right] & \frac{n}{2\sigma^4} - \frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma^6} \end{bmatrix} = -\frac{1}{n}\boldsymbol{I_n}(\boldsymbol{\theta}).$$

Now let $\boldsymbol{x_i}$ be the ith row of the design matrix $\boldsymbol{X}$, and y_i be the ith element of the observation vector y. With these constructs at hand, we can define the score components for individual observations in our sample as

$$\boldsymbol{U_i}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{(y_i - \boldsymbol{x_i}\boldsymbol{\beta})\boldsymbol{x_i}'}{\sigma^2} \\ \frac{-1}{2\sigma^2} + \frac{(y_i - \boldsymbol{x_i}\boldsymbol{\beta})^2}{2\sigma^4} \end{bmatrix}.$$

With these components defined, we can then use the preceding to represent the matrix $\boldsymbol{B_n}(\boldsymbol{\theta})$ used in the White estimator as

$$\boldsymbol{B_n}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U_i}(\boldsymbol{\theta})\boldsymbol{U_i}(\boldsymbol{\theta})'.$$

Thus, with $\boldsymbol{B_n}(\boldsymbol{\theta})$ and $\boldsymbol{A_n}(\boldsymbol{\theta})$ defined, these matrices can be evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ and be used to define

$$\begin{aligned} \boldsymbol{C_n}(\hat{\boldsymbol{\theta}}) &= A_n^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{B_n}(\hat{\boldsymbol{\theta}})A_n^{-1}(\hat{\boldsymbol{\theta}}) \\ &= n\boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}})\left[\sum_{i=1}^{n} \boldsymbol{U_i}(\hat{\boldsymbol{\theta}})\boldsymbol{U_i}(\hat{\boldsymbol{\theta}})'\right]\boldsymbol{I_n}^{-1}(\hat{\boldsymbol{\theta}}), \end{aligned}$$

where $\boldsymbol{C_n}(\hat{\boldsymbol{\theta}})$ will be a $(r+1)$ by $(r+1)$ matrix that can be used as an estimator of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$. This estimator is robust to misspecification of the model.

Consider the bottom rightmost element of this matrix, that being the $(r+1)^{th}$ element of the $(r+1)^{th}$ column of the matrix. This element will correspond to the large-sample robust variance of $\hat{\sigma}^2$. Let $s(\boldsymbol{\theta})$ refer to this corresponding element in the case of the theoretical matrix $\boldsymbol{C}(\boldsymbol{\theta})$, and $s_n(\hat{\boldsymbol{\theta}})$ refer to this corresponding element of the estimator $\boldsymbol{C_n}(\hat{\boldsymbol{\theta}})$. By White's results, it can then be seen that

$$\sqrt{n}(\hat{\sigma}^2 - \sigma_*^2) \xrightarrow{d} N(0, s(\boldsymbol{\theta}_*)), \tag{2}$$

where σ_*^2 denotes the pseudo-true parameter related to σ^2 in the case of potential misspecification.

We will again employ the delta method to transform the asymptotic distribution to a form that is more suitable. We once more define a transformation of

$$g(x) = \log(x).$$

Applying this transformation to the asymptotic distribution presented in (2), we arrive at the relation

$$\sqrt{n}(\log(\hat{\sigma}^2) - \log(\sigma_*^2)) \xrightarrow{d} N\left(0, \frac{1}{\sigma_*^4} s(\boldsymbol{\theta}_*)\right).$$

Using this asymptotic distribution, one can arrive at an approximate distribution for $n\log(\hat{\sigma}^2)$ as

$$n\log(\hat{\sigma}^2) \dot{\sim} N\left(n\log(\sigma_*^2), \frac{n}{\sigma_*^4} s(\boldsymbol{\theta}_*)\right),$$

which could be suitable for use in finite sample sizes that are sufficiently large. However, σ_*^2 and $s(\boldsymbol{\theta}_*)$ are unlikely to be unknown in practical modeling applications. Thus, estimating these quantities with $\hat{\sigma}^2$ and $s_n(\hat{\boldsymbol{\theta}})$ respectively, a reasonable estimate for the variance of $n\log(\hat{\sigma}^2)$ can be proposed as

$$Var\left[n\log(\hat{\sigma}^2)\right] \approx \frac{n}{\hat{\sigma}^4} s_n(\hat{\boldsymbol{\theta}}).$$

By the relation presented in (1), it is clear that this variance estimate is also suitable for $Var\left[-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]$, and therefore likelihood-based information criteria as a whole that possess a constant penalty term. This estimator should approximate our previously derived value $2n$ in the case of a correctly specified

model, as the sandwich estimator component will approximate the expected Fisher information used earlier, and the MLE $\hat{\sigma}^2$ should converge to the true parameter value σ^2. However, this sandwich estimator need not assume correct specification, and should be relatively robust to model misspecification. For the remainder of this work, this sandwich estimator will be referred to as $\widehat{Var}[GOF]$.

We have established an asymptotic variance for the likelihood goodness-of-fit term in the case of a correctly specified normal linear regression model, and an estimator for this variance that does not assume the model is correctly specified. We will now synthesize these two developments to form a general goodness-of-fit procedure to test the hypothesis that a given normal linear regression model is correctly specified.

Under this null hypothesis, the estimator $\widehat{Var}[GOF]$ should be close to the theoretical value $2n$ for a sufficient sample size. We propose the use of the non-parametric bootstrap to obtain an empirical estimate for the sampling distribution of $\widehat{Var}[GOF]$. Once this empirical distribution has been obtained, the null hypothesis can be tested by observing whether a bootstrap interval for $Var[GOF]$ contains the theoretical value $2n$, as the approximation $Var[GOF] \approx 2n$ should hold for sufficient sample sizes under the null hypothesis. If a $100 * (1 - \alpha)\%$ bootstrap confidence interval does not contain $2n$, we reject the null hypothesis and conclude that the model is misspecified; however, if the interval does contain $2n$, we do not have sufficient evidence to reject the null hypothesis, and the proposed model does not demonstrate lack-of-fit. A full summary of the proposed procedure can be found in the presented Algorithm.

This procedure can be used to assess the general hypothesis that a normal linear regression model is properly specified against the alternative that it displays lack-of-fit. Unlike other goodness-of-fit tests and procedures, this method does not test a specific assumption of linear regression such as normality or homoskedasticity, but rather many assumptions of normal linear regression. If some assumptions are violated, the property $Var[GOF] \approx 2n$ is not guaranteed to hold. This characteristic allows the test to detect many forms of misspecification from mean misspecification to distributional misspecification

Algorithm Bootstrap Goodness-of-Fit Test for a Normal Linear Regression Model

For test level α, candidate normal linear model M, bootstrap iterations B, sample size n, and a null hypothesis that M is adequately specified:

1: Resample, with replacement, outcomes with covariates to generate a bootstrap sample of size n.

2: Fit model M to this bootstrap sample, and with this fitted model, calculate $\widehat{Var}[GOF]$ and record this statistic.

3: Repeat steps 1-2 B times to generate an empirical bootstrap distribution for $\widehat{Var}[GOF]$.

4: Construct a $100 * (1 - \alpha)\%$ bootstrap confidence interval for $Var[GOF]$.

5: If this interval does not contain $2n$, reject the null hypothesis at the α level. If it does contain $2n$, the null hypothesis was not rejected and model M does not appear to exhibit lack-of-fit.

to heteroskedasticity.

Additionally, an implementation of this bootstrap test was developed in the R package `DBModelSelect` [9], available for download via CRAN and Github. The `BootGOFTestLM` function in this package provides a convenient and efficient way to perform the bootstrap test on a fitted linear model.

While the simulations in the following section will employ a percentile interval as the bootstrap confidence interval method of choice, one is not limited to using this method and may use any bootstrap interval they choose so long as it is theoretically justifiable. Additionally, the non-parametric bootstrap is employed in the algorithm detailed above to avoid further assumptions. Limited testing results suggest that the residual bootstrap may also work just as well in this procedure. However, pending further investigation, the non-parametric bootstrap is recommended at the current time. All simulations presented in this work will use the non-parametric bootstrap implementation of the procedure.

3. Simulation Studies

Five simulation settings will be employed to assess the efficacy of the bootstrap procedure developed in the previous section in comparison to the White test, Breusch-Pagan test, classical information matrix test, auxiliary regression information matrix test, and parametric bootstrap information matrix test. In each simulation, an n by 1 outcome vector $\boldsymbol{y} = [y_1, ..., y_n]'$ will be generated for $i = 1, ..., n$ according to

$$y_i = 2.0 + 2.0x_{i1} + 2.0x_{i2} + \epsilon_i,$$

with the nature of the error term ϵ and the fitted candidate model varying between different simulations. For all simulations, x_{i1} and x_{i2} are completely iid covariates generated according to a $Uniform(0, 5)$ distribution.

In the first simulation scenario, each ϵ_i will be generated as $\epsilon_i \overset{iid}{\sim} N(0, 4)$. With the generated data in hand, we will proceed to fit a normal linear regression model that includes an intercept and effects for x_{i1} and x_{i2}. This model is properly specified, and should not exhibit gross lack-of-fit.

In the second simulation scenario, once again each ϵ_i will be generated as $\epsilon_i \overset{iid}{\sim} N(0, 4)$. However, in this case the fitted model will be a normal linear regression model with an intercept and effect for only x_{i1}. The covariate x_{i2} will be omitted from the model, and could be viewed as a factor that affects the true generating process, but was unobserved in data collection. This omission will induce mean misspecification for this fitted model as the mean structure will be underspecified.

In the third simulation scenario, each ϵ_i will be generated as $\epsilon_i \overset{iid}{\sim} N(0, 4)$. However, we will generate a third variable x_{i3} such that $x_{i3} = 0.3x_{i2} + 0.7u_i$, where each u_i is drawn according to a $Uniform(0, 5)$ distribution. Then, the fitted model will be a normal linear regression model with an intercept and effects for x_{i1} and x_{i3}. This model is misspecified as while the covariate x_{i3} may serve as a surrogate for x_{i2} due to their relation, we are still missing information used in the true generating process.

In the fourth simulation scenario, each ϵ_i will be generated as

$$\epsilon_i \overset{iid}{\sim} N\left(0, (2 + x_{i3})^2\right),$$

and x_{i3} is an additional iid covariate generated according to a $Uniform(0, 5)$ distribution. Note that heteroskedasticity is introduced into this model courtesy of x_{i3} affecting the error variance. The fitted model for this simulation will be a normal linear regression model that has the correct mean structure of an intercept with effects for x_{i1} and x_{i2}, but will be fit assuming constant variance. This will result in unbiased regression parameter estimates; however, these estimates are not guaranteed to be the most efficient, and could be improved by including information regarding x_{i3}.

In the final simulation scenario, each ϵ_i will be generated as

$$\epsilon_i \overset{iid}{\sim} N\left(0, (2 + 0.5x_{i2})^2\right).$$

The fitted model for this simulation will be a normal linear regression model that has the correct mean structure of an intercept with effects for x_{i1} and x_{i2}, but will be fit assuming homoskedasticity. In this case the heteroskedasticity present is related to a covariate in the model matrix, and thus any normal linear model fit to the data will not be properly specified if it assumes homoskedasticity, leading to potentially improper inference.

In each simulation, the bootstrap goodness-of-fit test, White test, Breusch-Pagan test, classical information matrix test, auxiliary regression information matrix test, and parametric bootstrap information matrix test will each be performed at the $\alpha = .05$ level on the fitted model. We expect the tests to roughly maintain their Type I error rates in the cases where the null hypothesis of each is true, and reveal the power of the test in the simulations where the null hypothesis is violated. After many simulation iterations, the Type I error rates or power of each test will be calculated based on the proportion of times each test rejected its null hypothesis. Each simulation will be performed for $n = 50, 100, 250, 500, 750, 1000$ and 2500, with 1000 bootstrap iterations for all sets. The simulation will be generate 1000 samples for each value of n,

performing the tests each time.

The code for the simulations can be found at `https://github.com/shkoeneman/GOF_manuscript/tree/main/scripts`. The results for the first simulation can be found in Table 1.

Table 1: Simulation 1 Results - Type I Error

n	Bootstrap GOF	White	Breusch-Pagan	Classical IM	Regression IM	Bootstrap IM
50	0.038	0.041	0.042	0.598	0.766	0.027
100	0.088	0.033	0.051	0.511	0.622	0.032
250	0.081	0.055	0.050	0.383	0.430	0.031
500	0.078	0.050	0.057	0.238	0.260	0.034
750	0.082	0.046	0.038	0.204	0.218	0.025
1000	0.089	0.058	0.050	0.182	0.187	0.035
2500	0.061	0.036	0.046	0.095	0.098	0.031

The developed bootstrap test exhibits anti-conservative behavior across most values of n with the empirically determined Type I error rate exceeding the desired Type I error rate of $\alpha = 0.05$ in all scenarios. However, as n increases, the empirical Type I error rate appears to be decreasing towards the desired value, albeit somewhat slowly. Thus, when using the procedure, one should keep in mind that the test may be more prone to rejection than one might expect, particularly for small sample sizes before asymptotic properties of the test truly manifest. The parametric information matrix test shows mirrored behavior, possessing a conservative Type 1 error rate at all values of n. The classical and regression variants of the information matrix test fail to maintain a true Type 1 error rate close to the nominal value until the sample size is rather high, and thus should not be trusted outside of large sizes. The White and Breusch-Pagan tests roughly maintain the desired Type I error rate across all values of n.

The results for the second simulation can be found in Table 2.

Table 2: Simulation 2 Results - Power/Type I Error

n	Bootstrap GOF	White	Breusch-Pagan	Classical IM	Regression IM	Bootstrap IM
50	0.279	0.053	0.055	0.675	0.792	0.003
100	0.506	0.053	0.071	0.760	0.815	0.007
250	0.828	0.065	0.050	0.860	0.881	0.009
500	0.953	0.033	0.036	0.950	0.954	0.098
750	0.991	0.059	0.059	0.986	0.986	0.436
1000	0.996	0.055	0.048	0.993	0.995	0.730
2500	1.000	0.053	0.042	1.000	1.000	1.000

The power of the bootstrap goodness-of-fit test in this scenario of mean misspecification is rather modest for low values of n, but rises rapidly to high levels. Thus, the bootstrap test appears to possess the ability to detect this mean misspecification on account of an unobserved covariate, and the power to detect this misspecification rises as the sample size increases as one would expect. The parametric bootstrap variant of the information matrix test eventually reaches similar levels of power, but lags substantially behind until n increases over 1000. Thus, the bootstrap goodness-of-fit test appears to possess better power to detect this mean misspecification due to a missing covariate in sample sizes below the highest values employed in the simulation.

The classical and regression-based information matrix tests possess high power for all sample sizes, as is to be expected from their proclivity for rejecting null hypotheses even when they are true. This characteristic of the classical and regression information matrix tests will be present in the next simulation as well and will not be commented on as such.

In contrast, the White test and Breusch-Pagan test both seem to only maintain their Type I error rates established in Simulation 1. This performance was to be expected as the null hypotheses of these tests are not violated in this scenario. However, this simulation setting does demonstrate how the inability of a goodness-of-fit test to detect improper fit does not mean lack-of-fit is not present, as these two tests are limited in the scope of misspecification they are able to detect.

The results for the third simulation can be found in Table 3.

Table 3: Simulation 3 Results - Power/Type I Error

n	Bootstrap GOF	White	Breusch-Pagan	Classical IM	Regression IM	Bootstrap IM
50	0.138	0.040	0.039	0.934	0.972	0.005
100	0.275	0.035	0.027	0.951	0.970	0.005
250	0.498	0.058	0.040	0.988	0.991	0.049
500	0.654	0.052	0.035	0.998	0.999	0.486
750	0.792	0.059	0.027	0.999	0.999	0.873
1000	0.868	0.058	0.038	1.000	1.000	0.981
2500	0.996	0.115	0.036	1.000	1.000	1.000

The bootstrap goodness-of-fit test possesses higher power in the third simulation in small sample sizes than the bootstrap information matrix test. However, at $n = 750$ the bootstrap information matrix test overtakes the bootstrap goodness-of-fit test, and appears to hold a small lead in power until both tests level out at near perfect power for $n = 2500$. Thus, each test holds an advantage in power for different bands of sample sizes. The White and Breusch-Pagan tests once again seem to maintain their desired sizes fairly well.

We now move on to the fourth simulation. Figure 1 presents a residual plot generated by fitting the specified model for the fourth simulation setting to data generated as described for this setting. Note the seemingly patternless, constant variance spread, indicating that one would not find reason to think any assumption of linear regression was violated.

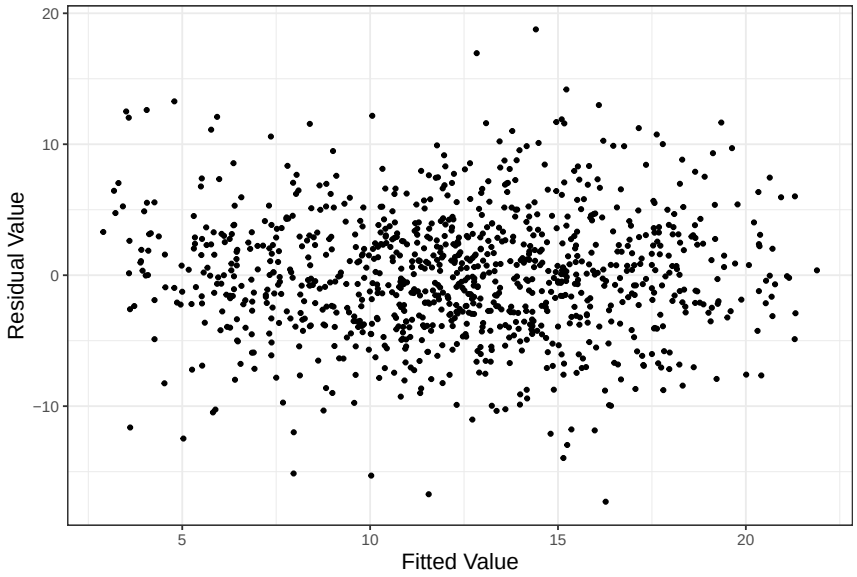

Figure 1: Simulation 4 Residual Plot.

The results for the fourth simulation can be found in Table 4.

Table 4: Simulation 4 Results - Power/Type I Error

n	Bootstrap GOF	White	Breusch-Pagan	Classical IM	Regression IM	Bootstrap IM
50	0.006	0.054	0.046	0.051	0.702	0.146
100	0.121	0.048	0.038	0.356	0.525	0.307
250	0.506	0.047	0.036	0.183	0.317	0.538
500	0.891	0.044	0.041	0.160	0.369	0.790
750	0.976	0.052	0.049	0.335	0.547	0.928
1000	0.994	0.062	0.055	0.547	0.692	0.972
2500	1.000	0.057	0.062	0.994	0.994	1.000

The bootstrap goodness-of-fit test appears to require a relatively high sample size to produce adequate power. However, the test does seem to eventually be able to detect this violation of homoskedasticity. The bootstrap information matrix test performs somewhat similarly, possessing higher power for small sample sizes and lower power for moderate sample sizes. The classic and regression variants of the information matrix test display erratic power as n increases, and thus further show their unreliability.

The White test and Breusch-Pagan test exhibit only power in line with their Type I error rates established earlier, and thus do not seem to be able to detect this form of heteroskedasticity, as the residual plot could not. This reveals how the subtleness of this homoskedasticity, which should not affect bias of estimates, but may affect efficiency.

This limitation in detection is on account of how the Breusch-Pagan and White tests and residual plots are constructed, as they rely on the detection of heteroskedasticity that is induced by observed covariates. When heteroskedasticity is induced by an unobserved covariate that is not used in the model, the White test and Breusch-Pagan test cannot ascertain that their null hypotheses of homoskedasticity are violated, and residual plots do not show any visible heteroskedasticity. However, the bootstrap goodness-of-fit test and variants of the information matrix test are able to reject their null hypotheses.

The results for the fifth and final simulation can be found in Table 5. Note that the null hypotheses of all tests are violated in this case, and each test should be able to detect this form of misspecification.

Table 5: Simulation 5 Results - Power

n	Bootstrap GOF	White	Breusch-Pagan	Classical IM	Regression IM	Bootstrap IM
50	0.018	0.169	0.327	0.725	0.875	0.146
100	0.036	0.314	0.695	0.855	0.912	0.307
250	0.141	0.773	0.992	0.988	0.992	0.538
500	0.413	0.995	1.000	1.000	1.000	0.790
750	0.583	1.000	1.000	1.000	1.000	0.928
1000	0.732	1.000	0.055	1.000	1.000	0.972
2500	0.991	1.000	1.000	1.000	1.000	1.000

In this scenario, the White test and Breusch-Pagan test both perform well in moderate to large sample sizes. This is likely due to the heteroskedasticity being induced by a covariate that is also in the proposed mean structure, the very case for which the White and Breusch-Pagan tests were designed. The bootstrap variant of the information matrix test performs similarly, while the classical and regression variants again display extreme power at low sample sizes. While the bootstrap goodness-of-fit test is able to detect this misspecification,

it requires a very large sample size to have adequate power.

In summary, we see that in general, the classical and regression variants of the information matrix test tend to reject null hypotheses no matter the circumstance, and fail to maintain acceptable type I error rates outside of extraordinary sample sizes. Thus, these tests are not recommended for general use. The White and Breusch-Pagan tests are the most adept at detecting heteroskedasticity related to covariates in the mean structure, but are narrow in their focus and cannot detect other forms of misspecification. The bootstrap information matrix test and bootstrap goodness-of-fit test can both detect myriad forms of misspecification, and both maintain acceptable type I error rates. The bootstrap information matrix test may possess better power in the presence of heteroskedasticity, and the bootstrap goodness-of-fit test may possess better power in settings where the mean structure is misspecified due to the omission of covariates. Thus, the new bootstrap goodness-of-fit test presents advantages over competetitor tests, particularly in its ability to detect multiple forms of misspecification and its ability to detect forms of mean misspecification where important covariates may be missing.

4. Discussion and Conclusion

In this work, we have developed a new bootstrap-based goodness-of-fit procedure to assess the goodness-of-fit of a fitted normal linear regression model. In order to develop this test, we first derived an asymptotic variance for the goodness-of-fit term present in likelihood-based information criteria such as AIC under the assumption that the fitted regression model is properly specified. The test functions by assessing whether a robust estimated variance conforms to the theoretical value under proper specification. Simulation studies demonstrated the ability of this bootstrap test to detect a wide variety of violations of assumptions inherent to normal linear regression. These violations include mean misspecification and violations to homoskedasticity. As such, this new procedure has the potential to serve as an omnibus goodness-of-fit procedure for normal

linear regression models. Additionally, we developed the `BootGOFTestLM` function of the R package `DBModelSelect` to provide an efficient and convenient method of performing the test.

This procedure has the potential to assist in model selection. For example, if one were to consider a normal linear regression framework to model a certain outcome, one may choose to employ the bootstrap test on the largest possible model in consideration, as if this model is misspecified, then all nested models will also be misspecified. If the test does not reject its null hypothesis of proper specification, one could proceed with the normal linear regression framework and select a final model using an information criterion such as Mallows' Cp, or by employing another method such as cross validation.

This bootstrap goodness-of-fit procedure is not without its drawbacks. While the test is able to detect heteroskedasticity induced by unobserved covariates, this may be viewed as a detriment, as this form of heteroskedasticity will not affect bias, and only efficiency. Without knowledge of unobserved covariates, one may not be able to improve efficiency. Thus, if the test rejects its null hypothesis but one still would like to move forward with a normal linear regression model, it is recommended to use robust estimates for the variance of estimated effects to hedge against performing improper inference. Simulations showed that the bootstrap test may be underpowered for small sample sizes. A bootstrap rendition of the information matrix test may possess better power to detect certain forms of misspecification. Additionally, the test may be overpowered for real-life datasets with large sample sizes, resulting in rejection of the null hypothesis when in fact a linear regression is a reasonable model. This is a feature common to many goodness-of-fit tests, and must be remembered when performing an analysis on observed data.

Potential future work involves finding similar procedures that can be used for other modeling frameworks.

5. Appendix

This Appendix will derive an exact variance for the goodness-of-fit term $-2\ell(\hat{\boldsymbol{\theta}})$ in the case of a properly specified linear model. While a similar derivation has been presented in other work [11], the following derivation differs substantially and is presented for completeness. We assume that a given linear regression model has a full rank n by r design matrix $\boldsymbol{X}$, outcome vector y, true parameters $\boldsymbol{\beta}$ and σ^2, and maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. We let

$$\boldsymbol{A} = \frac{1}{\sigma^2}(\boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$$

such that we can construct the quadratic form

$$\boldsymbol{y'Ay} = \frac{n\hat{\sigma}^2}{\sigma^2}.$$

By the properties of quadratic forms of multivariate normal random variables, it can be seen that

$$\boldsymbol{y'Ay} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}. \tag{3}$$

Note that if we take the natural logarithm of the left-hand side and evaluate its variance, we see that

$$Var\left[\log(\frac{n\hat{\sigma}^2}{\sigma^2})\right] = Var\left[\log\hat{\sigma}^2\right] \tag{4}$$

due to constant terms being irrelevant to the variance calculation. This enclosed form is very close to the $n\log\hat{\sigma}^2$ that is required to determine the variance of the likelihood goodness of fit term $-2\ell(\hat{\boldsymbol{\theta}})$ in the case of a normal linear regression model. Thus, if we can calculate $Var\left[\log\hat{\sigma}^2\right]$, which is itself the variance of the natural logarithm of a variable distributed as χ^2_{n-r}, we can easily obtain $Var\left[n\log\hat{\sigma}^2\right]$.

Let $W \sim \chi^2_\nu$, and let $Y = \log(W)$. Note that the moment generating function of Y can be expressed as

$$M_Y(t) = E\left[e^{yt}\right] = E\left[e^{t\log(w)}\right] = E\left[e^{\log w^t}\right] = E\left[w^t\right].$$

Thus, the moment generating function of Y evaluated at t is the t^{th} moment of W. The moments of W will have the closed form of

$$E\left[w^t\right] = 2^t \frac{\Gamma(t+\frac{\nu}{2})}{\Gamma(\frac{\nu}{2})} = E\left[e^{yt}\right] = M_Y(t).$$

Taking the first derivative with respect to t of the moment generating function $M_Y(t)$ and evaluating at $t = 0$, we find the first moment of Y to be

$$E\left[Y^1\right] = \log 2 + \psi^{(0)}\left(\frac{\nu}{2}\right),$$

where $\psi^{(0)}(z)$ is the digamma function. Thus, we have a form for the first moment of the logarithm of a central chi-squared random variable. By taking another derivative with respect to t and evaluating at $t = 0$, we find the second moment to be

$$E\left[Y^2\right] = (\log 2)^2 + \psi^{(0)}\left(\frac{\nu}{2}\right) 2\log 2 + \psi^{(1)}\left(\frac{\nu}{2}\right) + \left(\psi^{(0)}\left(\frac{\nu}{2}\right)\right)^2,$$

where $\psi^{(1)}(z)$ is the digamma function.

Therefore, the variance of Y, the log of a central chi-squared random variable with ν degrees of freedom, is found to be

$$Var\left[Y\right] = E\left[Y^2\right] - \left(E\left[Y\right]\right)^2 = \psi^{(1)}\left(\frac{\nu}{2}\right).$$

This value can be calculated using software to approximate the trigamma function, although it should be noted that this calculation may be unstable for certain values of the degrees of freedom.

Recalling the distributional result established in (3) and the relation in (4), it is clear that

$$Var\left[n\log\hat{\sigma}^2\right] = Var\left[-2\ell(\hat{\boldsymbol{\theta}})\right] = n^2\psi^{(1)}\left(\frac{n-r}{2}\right).$$

However, while this variance is exact, it still involves an approximation as the trigamma function must be approximated. Additionally, it was found that this variance provided no benefit when used in the bootstrap goodness-of-fit procedure developed in this work. Thus, this derivation is presented here only for completeness and as a matter of theoretical interest.

References

[1] Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] Breusch, T.S., and Pagan, A.R. (1989). A Simple Test for Heteroskedasticity and Random Coefficient Variation. *Econometrica*, **47**, 1287–1294.

[3] Cavanaugh, J.E., and Neath, A.A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Comput Stat*, 11:e1460.

[4] Chesher, A. (1983). The Information Matrix Test: Simplified Calculation via a Score Test Interpretation. *Economics Letters*, **13**, 45–48.

[5] Dhaene, G., and Hoorelbeke, D. (2004). The information matrix test with bootstrap-based covariance matrix estimation. *Economics Letters*, **82**, 341–347.

[6] Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. R. Soc. Lond. A*, **222**, 309–368.

[7] Freedman, D.A. (2006). On The So-Called 'Huber Sandwich Estimator' and 'Robust Standard Errors'. *The American Statistician*, **60**, 299–302.

[8] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **5**, 221–233.

[9] Koeneman, S.H. (2023). *DBModelSelect: Distribution-Based Model Selection*. R package version 0.2.0, `https://CRAN.R-project.org/package=DBModelSelect`.

[10] Kutner, M.H., Nachtsheim, C.R., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin, New York.

[11] McQuarrie, A.D.R., and Tsai, C-L (1998). *Regression and Time Series Model Selection.* World Scientific, New Jersey.

[12] Miles, J. (2014). Residual plot. *Wiley StatsRef: Statistics Reference Online.*

[13] Millar, R.B. (2011). *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB.* Wiley, Hoboken.

[14] Rao, C.R. (1965). *Linear Statistical Inference and Its Applications.* Wiley, Hoboken.

[15] Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **22**, 461–464.

[16] White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48**, 817–838.

[17] White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, **50**, 1–25.

[18] Waldman, D.M. (1983). A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity. *Economics Letters*, **13**, 197–200.