A Comparison of WK3and MSE for Regression Model Fitting

Wasfi Taher Kahwachi[1]
Tishik International University, Erbil, Iraq
Email: wasfi,kahwachit@ishik.edu.iq

Aras J. Mhamad[2]
Statistic Department, School of Administration & Economics, Sulaymaniyah University,
Sulaymaniyah, Iraq

DOI: 10.23918/ICABEP2019p9

Abstract

Wasfi Kahwachi3 (WK3) measurement is considered as one of the powerful statistical tools to testing goodness of fit of regression model when compared with the MSE (Mean Squared Error) measure. In this paper, the main goal is to detect the best measure for model fitting by using a comparison between MSE and WK3 measure. This new measure (WK3) applied on five samples to compare its value with MSE value which obtained from the same samples to judge on the new criterion for fitting the models. The results showed that the WK3 measure was better (smaller) than the MSE measure for fitting regression model in all samples. This gives the conclusion that Wk3 can be used besides MSE for testing goodness of fit in regression analysis.

Keywords: WK3 Measure, MSE Measure, Comparison MSE and WK3

1. Introduction

Values of MSE may be used for comparative purposes. Two or more statistical models may be compared using their MSEs as a measure of how well they explain a given set of observations. An unbiased estimator, based on a statistical model, with the smallest variance among all unbiased estimators is the best prediction in that it minimizes the variance and is called the minimum variance unbiased estimator [2] [4].

2. Materials and Methods

2.1 Goodness of fit analysis

After fitting a linear model relating Y to X, it is interested not only in knowing whether a linear relationship exits, but also in measuring the quality of the fit of the model to the data [3]. Some methods of goodness of fit analysis are employed here. A key measure of the strength of the association is the MSE, which is basically the average of the squared residuals. MSE is a measurement that gives an indication of the accuracy of the model [1][11].

If this value is small with respect to the sample variance of the y's, then we consider our regression model to be a worthwhile explanation of the association, this reflects into the value of F test.

2.2 Interpretation MSE in Regression

In regression analysis, the term mean squared error is sometimes used to refer to the error variance. Note that, although the MSE is not an unbiased estimator of the error variance, it is consistent, given the consistency of the predictor [10]. It is often referred to as mean squared prediction error or "out-of-sample mean squared error". An MSE of zero, meaning that the estimators predict observations of the parameter with perfect accuracy, is the ideal, but is practically never possible [5] [8].

In the regression techniques, analysis of variance estimate the MSE as part of the analysis and use the estimated MSE to determine the statistical significance of the factors or predictors under study. Thus, small values of MSE may refer to better fitting a certain set of data. [2].

2.3 Wasfi Kahwachi (WK3) Measurement

Wasfi Kahwachi* has developed the following criterion to measure the error behavior. The function gives us an idea about the behavior of the error occurring from a model fitting. It can be seen from the application that it could be used to study the error behavior like MSE. Big values of MSE are giving big values for WK3. This indicates that both of them having the same direction. It studies the squared error observations ratio divided by its number, resulting in a criterion that measures the average of the squared error ratios.

The question that raises here is it possible to be used besides MSE? The function of WK3 is given by:

$$WK3 = \sum_{i=1}^{n} \frac{ei^2+1}{ei^2} /(n-1) \qquad \ldots . \qquad (1)$$

The statistical distribution of *WK3* is supposed to be F, in an assumption of the independency, as in the appendix.

3. Data Analysis and Results

The data set in the study is used in order to compare and study the performance of the proposed criterion (*MSE* vs. *WK3*). Five data sets of different sample sizes were used aiming to study (*MSE* vs. *WK3*). Samples sizes were n = 100, 42, 30, 50 and 15. Regression models were used to fit the five data sets as follows:

The first data set consisted of a sample of 100 observations (patients) and was obtained from center of Diabetes in Sulaimaniyh governorate, the variables which are determined as explanatory variables were; Gender, Age, Rate of Cholesterol, Rate of Triglycerides, with response variable Rate of Sugar.

The second data set consisted of a sample of 42 observations and was obtained from Statistics Bureau in Sulaimaniyh governorate. The explanatory variables were; yield Kg/Acre (X1), Area / Acre (x2), Rain/ml (x3), with response variable Production (y).

The third data set consisted of a sample of 30 observations (patients), the explanatory variables were; AGE (x1), Weight (x2), Parents (x3), Smoke (x4), Exercise (x5), with response variable Systolic (y).

The fourth data set consisted of a sample of 50 observations[1], the explanatory variables were; reported violent crime rate per 100,000 residents (X2), annual police funding in $/resident (x3), % of people 25 years+ with 4 yrs. of high school (x4), % of 16 to 19 year-olds not in high school and not high school graduates (x5), % of 18 to 24 year-olds in college (x6), % of people 25 years+ with at least 4 years of college (x7), with response variable total overall reported crime rate per 1 million residents (y).

The fifth data set consisted of a sample of 15 observations and was obtained from Statistics Bureau, the College of Agriculture, and the Ministry of Agriculture – Iraq. The explanatory variables were; the amount of yielding import production of the province (x1), the trefoil amount of yielding production of the province (x2), dust amount of yielding production of the province (x3), corn amount of yielding production of the province (x4), feed mixture amount of yielding production of the province (x5), the amount of humidity in the provinces (x6), the amount of rainfall in the province (x7), Temperatures in the province (x8), with response variable; Number of cows (y).

Finally, it is obvious that the *WK3* measure is less than the *MSE* measure in the five samples as in table (1):

Table (1): Comparison between the values of *MSE* with *WK3*

Sample	*MSE*	*WK3*	Decision
Sample 1	3548.951402	104.0028055	*MSE* > *WK3*
Sample 2	85426183.04	779.1158207	*MSE* > *WK3*
Sample 3	11.81196146	3.250819038	*MSE* > *WK3*
Sample 4	38086.57867	66.82857057	*MSE* > *WK3*
Sample 5	3620710361	57.55427025	*MSE* > *WK3*

4. Conclusions (Discussion)

In this study, the methodology provided a powerful criterion supporting testing goodness of fit models by comparing between *MSE* and *WK3* (developed by Wasfi Kahwachi). The main aim for the conducted study was to study the behavior and comparison between *MSE* and *WK3* values. This new measure *WK3* applied on the five samples for model fitting of regression model. The result showed that the *WK3* measure is smaller than the *MSE* measure, the value of *MSE* for the samples were (3548, 85426183, 11, 38086, and 3620710361) respectively, and the *WK3* for the samples were (104, 779, 3, 66, and 57) respectively. It's clear that the *KW3* was always less than the *MSE* for the regression analysis.

5. References

Cavanaugh E. Joseph(2012)."Model Selection: Criteria for Regression Model Selection" Department of Biostatistics, Department of Statistics and Actuarial Science, the University of Iowa

Chattefuee Samprit & Hadi Ali S .(2006)."Regression Analysis by Example". Fourth Edition, Wiley Series in Probability and Statistics.

Geyer J. Charles (2002)." Model Selection in R". (Publishing company)

Gilroy J. E. , Hirsch M. R. & Cohn A. T. (1990)."Mean Square Error of Regression - Based Constituent Transport Estimates". Water Resources, VOL. 26, NO. 9.

Huddleston F. Harold (1997)."A Theory of Minimum Mean Square Estimation in Surveys with Non response". Statistical Reporting Service, U.S. Department of Agriculture.

Jr Harrell E Frank (2004). "Biostatistical Modeling". Department of Biostatistics, Vanderbilt University School of Medicine. Nashville TN USA.

Liski P. Erkki, Toutenburg Helge & Trenkler Gijtz (1993)." Minimum mean square error estimation in linear regression". Journal of Statistical Planning and Inference 37, 203-214, North-Holland.

Mohammed A.AL-Fawzan (2000). "Algorithms for Estimating the Parameters of the Weibull Distribution", King Abdul-Aziz City for Science and Technology, Riyadh, Saudi Arabia.

Orlov L. Michael (1996)." Multiple Linear Regression Analysis Using Microsoft Excel ". Chemistry Department, Oregon State University.

Schunn, Christian (2002)."Evaluating Goodness-of-Fit in Comparison of Models to Data", University of Applied Sciences Kaiserslautern, University of Pittsburgh, USA.

Zaka Azam & Akhter Saeed Ahmad (2005)."Methods for Estimating the Parameters of the Power Function Distribution ". College of Statistical and Actuarial Sciencess University of the Punjab, Lahore.

Appendix

Here we want to find the distribution of *WK3*.

Since $e_i \sim$ IIN(0,σ^2),

Hence both e_i^2 and e_{i+1}^2 are distributed independently as χ^2 and (e_i $assumption$), then *WK3* given by:

$WK3 = \sum_{i=1}^{n} \frac{ei^2+1}{ei^2}/(n-1)$ is distributed as F.

Proved