# Model Selection by Friedman Statistic

Adil Korkmaz
Akdeniz University Faculty of
Economics and Social Sciences
Antalya, 07058 Turkey
adilkorkmaz@akdeniz.edu.tr

M. Burak Önemli
Kansas State University
Department of Statistics
Dickens Hall, Manhattan
KS, 66506
mbonemli@k-state.edu

## Abstract

This study investigates an application of Friedman test statistic as a model selection methodology on post estimation data. Although there are various model selection criteria, their main focus is to fit the model to the estimation data. Some of these criteria are appropriate for nested model selection while the rest is suitable for non-nested model selection. The suggested model selection methodology is indifferent to the distinction between nested or non-nested model selections. In suggested methodology, the dataset are sub divided into two parts: First part is used to estimate all the competing models while the rest are used for performance comparison of competing models. In this respect, all competing models are ranked according to their forecast performance based on proximity between observed and estimated values of the dependent variable in post estimation data. Then, we showed that the suitability of Friedman test statistic in order to evaluate the prediction performance of competing models.

**Keywords**: Model Selection, Econometric Model, Friedman Statistic, Prediction Performance, Post Estimation Data.
2000 AMS Classification: 60F05

## 1. Introduction

Model selection among many competing models is one of the foremost topics in regression analysis. The importance of model selection emerges because an inappropriate model specification results in serious problems including biased estimates due to an omission of relevant variables and inefficient estimates due to inclusion of irrelevant variables. Therefore, discovering the best model among many competing models lies at the core of model specification. The best model search is a two-stage process. In the first stage, potentially important factors or predictors are determined with consideration for unbiased and efficient estimates. In the second stage, subset models that can be formed from predictors are evaluated and competed. From any set of $K$ predictors, the number of alternative models that can be constructed is $2^K$ (Kutner et al., 2005). It is noteworthy that there is a trade-off between parsimony and precision, and model selection based on estimation data is a procedure for seeking an optimum. In other words, model selection based on estimation data could be evaluated as an optimality search between goodness of fit and parsimony. There are many

proposed methods for model selection. Some of these methods are appropriate for nested model selection while the rest are suitable for non-nested model selection. The common property of both nested and non-nested model selection methodologies is to focus on goodness-of-fit to estimation data. However, the suggested methodology is applicable for post estimation data. Furthermore, the suggested methodology is appropriate for both nested and non-nested model selections since the focus is goodness-of-fit of predictions to observations in post estimation data. Therefore, the suggested methodology is indifferent to the distinction between the nested and non-nested model selections.

Perhaps the most famous and naïve criterion for nested model selection is the determination coefficient, denoted by $R^2 = 1 - RSS/TSS$ where $RSS$ is residual sum of squares and $TSS$ is total sum of squares. However, due to a number of serious problems associated with $R^2$, its usage is not advised in general Asteriou (2006). The major criticism of $R^2$ is that due to the fact that the addition of an explanatory variable cannot cause this statistic to fall. Therefore, $R^2$ would always lead one in favor of larger model. Note that in general, adding a variable to the model increases the precision but decreases the parsimony, vice versa. An alternative criterion for model selection is adjusted $R^2$. Adjusted $R^2$, denoted by $\overline{R}^2$, is obtained by correcting the $R^2$ statistic for degrees of freedom. $\overline{R}^2$ is given by $1-(1-R^2)(T-1)/(T-K)$ where T is the number of observations used for estimation, and K is the number of parameters to be estimated. This criterion is preferable to $R^2$ since inclusion of an irrelevant variable is limited.

Some of the other criteria are: Akaike information criterion (AIC), which minimizes $ln(SSR/T)+2K/T$, Shwarz Bayesian information criterion SBIC, which minimizes $(SSR/T)+(K\,lnT)/T$, Amemiya's prediction criterion, which minimizes $SSR(1+K/T)/(T-K)$, and Hannan and Quinn criterion that minimizes $T\,ln(RSS/T)+2K\,ln(T)$ (Akaike 1974, Shwarz 1978, Hannan and Quinn 1979, Amemiya 1980, Judge 1985, Kennedy 2003). Note that these criteria are mainly used for non-nested model selection. Other criteria include Mallow's C criterion and prediction sum of squares (PRESS) criterion (Kutner et al., 2005). Each criterion gives different weights for the trade-off between parsimony and precision.

In addition, there are other methods called automatic search procedures. These methods are especially useful when the number of predictors is large, and therefore, all possible models require numerous trials. Note that automatic search procedures are not necessarily competitors of the model selection criteria. Conversely, when there are too many covariates, the model selection criteria are used with automatic search procedures. The automatic search procedures are variety of automatic computer based search procedures and mainly appropriate for nested model selection (Kutner et al., 2005). The mostly used automatic search procedures are: Best Subsets Algorithms and Stepwise Regression Methods.

In this study, we propose use of the Friedman test for model selection by focusing on post estimation analysis of competing models. With this motivation, the rest of this study is organized as follows. Section 2 discusses measuring prediction performances of competing models. The Friedman test statistic for model selection requires a series of hypothesis testing procedures. These procedures are derived in Section 3. Section 4 concludes.

## 2. Measuring Prediction Performances

We assume that there are $c$ competing models where $c > 1$. Furthermore, we assume that the total observations are divided into two groups. The first $m$ observations are the estimation data and the remaining $n$ observations are the post estimation data. Suppose Model 1, Model 2, …, Model $c$ are estimated based on the estimation data. Then, based on the post estimation data, the prediction performances can be measured as follows. Let $y_1$, $y_2$, …, $y_n$ denote the observations of the dependent variable in the post estimation data. Furthermore, let $\hat{y}_{1j}$, $\hat{y}_{2j}$, …, $\hat{y}_{nj}$ where $j \in \{1, 2, \ldots, c\}$ denote the predicted values of the dependent variable from each model. Hence, the prediction performances can be measured by smallness of the distances, defined by

$$a_{ij} = | y_i - \hat{y}_{ij} |$$

where $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, c\}$. It is instructive to note that the prediction performance is measured separately for each observation of the post estimation data. Nevertheless, it is also noteworthy to state that the prediction performance may be measured by smallness of any positive power transformation of the distance. However, this does not affect the order of the competing models because this order is invariant to such a transformation.

It is assumed that the dependent variable observations $y_1$, $y_2$, …, $y_n$ are independent of each other in regression analysis. On the other hand, the predictions $\hat{y}_{1j}$, $\hat{y}_{2j}$, …, $\hat{y}_{nj}$ are assumed to be constants since they are simply produced from the competing models and the post estimation data of independent variables. Related to this, $a_{ij}$'$s$ are independent of each other. To see this, consider two random variables $U$ and $V$. If $U$ and $V$ are independent of each other and also if $u$ and $v$ are constants, then $U - u$ and $V - v$ are also independent of each other. The distances $a_{i1}$, $a_{i2}$, …, $a_{ic}$ where $i \in \{1, 2, \ldots, n\}$ play a crucial role in determining model rankings. Generally speaking, a smaller rank must be assigned to a model with a relatively small distance. Then, the preference data are obtained by the formula:

$$R_{ij} = \text{the rank of } a_{ij} \text{ in the set of } a_{i1}, a_{i2}, \ldots, a_{ic} \text{ in an ascending order}$$

for $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, c\}$. Hence, the preference data is obtained as follows:

**Table 1: Preference Data**

| | **Model 1** | **Model 2** | **…** | **Model c** |
|---|---|---|---|---|
| **Observation (Block) 1** | $R_{11}$ | $R_{21}$ | … | $R_{1c}$ |
| **Observation (Block) 2** | $R_{21}$ | $R_{22}$ | … | $R_{2c}$ |
| **…** | … | … | … | … |
| **Observation (Block)** $n$ | $R_{n1}$ | $R_{n2}$ | … | $R_{nc}$ |
| **Total** | $R_{.1}$ | $R_{.2}$ | … | $R_{.c}$ |

It is obvious that the set of the $i$ th row ranks $R_{i1}$, $R_{i2}$, …, $R_{ic}$ is a permutation of the integers 1, 2, …, $c$ for every $i \in \{1, 2, \ldots, n\}$ and the integers from 1 to $c$ are assigned as rankings from the smallest distance to the largest for the competing models.

## 3. The Null Hypothesis and the Test Statistic

The null hypothesis is that there is no winner in the competing models. If there is no winner in the competing models, then one can assign any permutation of 1, 2, …, $c$ equally likely as the ranks to the competing models. Hence, under the null hypothesis,

$$P\left(R_{i1} = k_1, R_{i2} = k_2, \ldots, R_{ij} = k_j, \ldots, R_{ic} = k_c\right) = \frac{1}{c!} \tag{1}$$

where $k_1$, $k_2$, …, $k_c$ are any permutation of 1, 2, …, $c$ for every $i \in \{1, 2, \ldots, n\}$. It follows from (1) that the marginal probability function of $R_{ij}$ is given by

$$P(R_{ij} = k_j) = \frac{1}{c} \tag{2}$$

where $i \in \{1, 2, \ldots, n\}$, $j \in \{1, 2, \ldots, c\}$ and $k_j \in \{1, 2, \ldots, c\}$. This result can be proven as follows: Consider that the marginal probability function $P\left(R_{ij} = k_j\right)$ is equal to the sum of all the probabilities

$$P\left(R_{i1} = k_1, R_{i2} = k_2, \ldots, R_{ij} = k_j, \ldots, R_{ic} = k_c\right)$$

where $k_1$, $k_2$, …, $k_j$, …, $k_c$ are any permutation of 1, 2, …, $c$ under the condition that $R_{ij}$ is fixed as $k_j$. If $R_{ij}$ is fixed as $k_j$, then there are $(c-1)!$ equally likely permutations for $k_1$, $k_2$, …, $k_j$, …, $k_c$. Then

$$P(R_{ij} = k_j) = (c-1)!\frac{1}{c!} = \frac{1}{c}.$$

Notice that the conditional probability of the event $R_{ij_1} = k_{j_1}$ under the condition of the event $R_{ij_2} = k_{j_2}$ is given by:

$$P(R_{ij_1} = k_{j_1} \mid R_{ij_2} = k_{j_2}) = \frac{1}{c-1} \tag{3}$$

where $j_1 \neq j_2 \in \{1, 2, \ldots, c\}$. The proof from (2) to (3) is similar to that from (1) to (2).

**Theorem 1:** Let $R_{.j}$ and $u=\begin{pmatrix}U_1 & U_2 & \cdots & U_c\end{pmatrix}_{1\times c}$ are defined by $R_{.j}=\sum_{i=1}^{n}R_{ij}$ and $u'=\sqrt{\frac{12}{nc(c+1)}}\left(R_{.1}-\frac{n(c+1)}{2} \quad R_{.2}-\frac{n(c+1)}{2} \quad \cdots \quad R_{.c}-\frac{n(c+1)}{2}\right)'_{1\times c}$, respectively. Then, $E(u')=\begin{pmatrix}0 & 0 & \cdots & 0\end{pmatrix}_{1\times c}$.

**Proof:** Note that

$$E(R_{ij})=1\cdot\frac{1}{c}+2\cdot\frac{1}{c}+\cdots+c\cdot\frac{1}{c}=\frac{c+1}{2} \tag{4}$$

by equation (2). Furthermore, the expected value of $R_{.j}$ is given by:

$$E(R_{.j})=E(R_{1j})+E(R_{2j})+\cdots+E(R_{nj}). \tag{5}$$

As a consequence of (4) and (5), we have

$$E(R_{.j})=\frac{n(c+1)}{2} \tag{6}$$

where $j\in\{1,2,\ldots,c\}$. Thus, equation (6) completes the proof of Theorem 1.

**Theorem 2:** $Var(u)=I-\frac{1\cdot 1'}{c}$ where $I$ is $c\times c$ unit matrix and $1=\begin{pmatrix}1 & 1 & \cdots & 1\end{pmatrix}'_{1\times c}$.

**Proof:** Notice that

$$E(R_{ij}^2)=1^2\cdot\frac{1}{c}+2^2\frac{1}{c}+\cdots+c^2\frac{1}{c}=\frac{(c+1)(2c+1)}{6} \tag{7}$$

by equation (2). It follows from (4) and (7) that

$$Var(R_{ij})=E(R_{ij}^2)-(E(R_{ij}))^2=\frac{c^2-1}{12}. \tag{8}$$

Since the observations (distances) are assumed independent of each other, $Var(R_{.j})$ is:

$$Var(R_{.j})=\frac{n(c^2-1)}{12}. \tag{9}$$

On the other hand,

$$E(R_{ij_1}R_{ij_2})=\sum_{\ell_1\neq\ell_2=1}^{c}\ell_1\ell_2\frac{1}{c}\cdot\frac{1}{c-1} \tag{10}$$

because $P(R_{ij_1}=\ell_1, R_{ij_2}=\ell_2)=P(R_{ij_1}=\ell_1)\cdot P(R_{ij_2}=\ell_2 \mid R_{ij_1}=\ell_1)=\frac{1}{c}\frac{1}{c-1}\cdot$ Notice that $\sum_{\ell_1\neq\ell_2=1}^{c}\ell_1\ell_2=\left(\sum_{\ell=1}^{c}\ell\right)^2-\sum_{\ell=1}^{c}\ell^2$. Then, it follows from (10) and (4) that

$$Cov\left(R_{ij_1},R_{ij_2}\right)=E\left(R_{ij_1}R_{ij_2}\right)-E\left(R_{ij_1}\right)E\left(R_{ij_2}\right)=-\frac{c+1}{12}\cdot \tag{11}$$

Moreover, distances independency implies that:

$$Cov\left(R_{\cdot j_1},R_{\cdot j_2}\right)=-\frac{n(c+1)}{12}\cdot \tag{12}$$

By using definition of $u$, $Var(u)$ is given by:

$$\frac{12}{c(c+1)}\cdot Var\left(\left(R_{\cdot 1}-\frac{n(c+1)}{2}\quad R_{\cdot 2}-\frac{n(c+1)}{2}\quad\cdots\quad R_{\cdot c}-\frac{n(c+1)}{2}\right)'\right) \tag{13}$$

Substituting both (9) and (12) into (13) yields:

$$Var(u)=\begin{pmatrix} 1-\frac{1}{c} & -\frac{1}{c} & \cdots & -\frac{1}{c} \\ -\frac{1}{c} & 1-\frac{1}{c} & \cdots & -\frac{1}{c} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{c} & -\frac{1}{c} & \cdots & 1-\frac{1}{c} \end{pmatrix}_{c\times c}. \tag{14}$$

This completes the proof of Theorem 2.

**Lemma 1:** $\frac{R_{\cdot j}-E\left(R_{\cdot j}\right)}{\sqrt{Var\left(R_{\cdot j}\right)}}\sim N(0,1)$ as $n\to\infty$ for each $j\in\{1,2,\ldots c\}$.

**Proof:** By equations (4) and (8), the components of $R_{\cdot j}$ have finite means and finite variances. These conditions are sufficient (even if not necessary) for the central limit theorem (Dudewicz and Mishra, 1988). This completes the proof of Lemma 1.

Now, we can introduce the test statistic for testing the null hypothesis. To do this, consider the following quadratic form $Q=u'\left(I-\frac{11'}{c}\right)u$. It is possible to show that this quadratic form is equivalent to the Friedman statistic, and it is appropriate for testing the null hypothesis. To show this, first the following lemma is presented:

**Lemma 2:** $1'u = 0.$

**Proof:** Notice that $\sum_{j=1}^{c} R_{.j} = \frac{nc(c+1)}{2}\cdot$ On the other hand, $1'u$ can be shown to be equal to $\sqrt{\frac{12}{nc(c+1)}}\cdot\left(\sum_{j=1}^{c} R_{.j} - \sum_{j=1}^{c}\frac{n(c+1)}{2}\right)\cdot$ Therefore, substituting the former into the latter completes the proof of Lemma 2.

Hence, as a result of Lemma 2, $Q$ can be rewritten as follows:

$$Q = u'u, \tag{15}$$

and its distribution is provided in following theorem.

**Theorem 3:** $Q \sim \chi^2_{c-1}$ as $n \to \infty$.

**Proof:** Since $I - \frac{1\cdot 1'}{c}$ is an idempotent matrix, its eigenvalues $(\lambda_i's)$ are equal to either 0 or 1 with the number of 1's is equal to its trace, which is $c-1.$ Hence,

$$P'\left(I - \frac{1\cdot 1'}{c}\right)P = Diag\begin{pmatrix}1 & 1 & \cdots & 1 & 0\end{pmatrix} = Diag\left(\lambda_i\right) \tag{16}$$

where $P_{c\times c}$ is an orthogonal matrix with $P'P = PP' = I$ and $Diag\left(\lambda_i\right)$ is a $(c\times c)$-dimensional diagonal matrix. Therefore, it follows that

$$\left(I - \frac{1\cdot 1'}{c}\right) = PDiag\left(\lambda_i\right)P'. \tag{17}$$

In addition, let $v = \begin{pmatrix}V_1 & V_2 & \cdots & V_c\end{pmatrix}'_{1\times c}$ be defined by $v = P'u$. By using equation (6), one can obtain $E(v) = \begin{pmatrix}0 & 0 & \cdots & 0\end{pmatrix}'$. Similarly, $Var(v) = Diag\left(\lambda_i\right)$ can be derived from equation (16). Hence, $V_1$, $V_2$, …, $V_{c-1}$ are standardized random variables since their mean values are equal to 0 and their variances are equal to 1. In addition, the degenerate random variable $Z_c = 0$ since both its mean value and variance are all equal to 0. Further, $Var(v)$ also denotes that the covariance between any pair of $V_1$, $V_2$, …, $V_{c-1}$ is equal to 0 while $V_1$, $V_2$, …, $V_{c-1}$ have a limiting $N(0,1)$ distribution by virtue of both Lemma 1 and the definition of $v = P'u$. These conclusions imply that $V_1$, $V_2$, …, $V_{c-1}$ are mutually independent when $n \to \infty$. Hence, by the definition of $Q$ given in (15) can be written as

$$Q = u'u = v'PP'v = v'v = V_1^2 + V_2^2 + \cdots + V_{c-1}^2. \tag{18}$$

On the other hand, since $V_1$, $V_2$, …, $V_{c-1} \sim N(0,1)$ and mutually independent as $n \to \infty$, $Q$ has a limiting chi-square distribution with $c-1$ degrees of freedom, which completes the proof of Theorem 3.

Notice that by equation (15), $Q$ is shown to be equivalent to the following:

$$Q = \frac{12}{nc(c+1)} \cdot \sum_{j=1}^{c} \left( R_{.j} - \frac{n(c+1)}{2} \right)^2 = \frac{12}{nc(c+1)} \cdot \sum_{j=1}^{c} R_{.j}^2 - 3n(c+1). \qquad (19)$$

The statistic given in equation (19) is obviously the Friedman statistic having a limiting chi-square distribution with $c-1$ degrees of freedom as proven in Theorem 3. Therefore, the rule of testing hypothesis when $n$ is large is given by

Reject $H_0$ if $Q > \chi^2_{c-1,\alpha}$; otherwise do not reject,

where $\alpha$ is the first type error and $\chi^2_{c-1,\alpha}$ is the upper $\alpha$ percentile point of a chi-square distribution with $c-1$ degrees of freedom. Notice that rejecting the null hypothesis based on the above rule requires rejecting validity of equation given in (1), which implies that at least some of the competing models have better prediction performances than the others have. To discover the 'great' winner of all the competing models, the above procedure should be repeated by eliminating the ‘weakest’ model, to which the largest rank mostly assigned.

## 4. Conclusion

The primary objective of this paper is to introduce a methodology that is based on the Friedman statistic for model selection. The suggested procedure depends on the prediction performances that can be measured distances on post estimation data. We showed that the prediction performances can be used to rank models, and the test statistic calculated from ranked models is equivalent to the Friedman test statistic. The procedure suggested here can be an alternative for other commonly employed model selection criteria. The main advantage of this methodology it is appropriate for both nested and non-nested model selection since it compares prediction performances. Given that no attempt is made to measure performance of Friedman test statistic in model selection, then carrying out some empirical comparison of this procedure with other criteria and identifying differences in how each criterion handles the statistical priorities of model selection would prove fruitful in terms of future research.

## Acknowledgements

The authors wish to thank Robert Poulson and Heath Yates and two anonymous referees for helpful comments and suggestions.

## References


1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control,* 19 (6), 716-723.
2. Amemiya, T. (1980). Selection of regressors. *International economic review,* 21, 331-354.
3. Asteriou, D. (2006). *Applied econometrics*. New York. Palgrave Macmillan.
4. Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association,* 32, 675-701.
5. Graybill, F. A. (1961). *An introduction to linear statistical models.* New York. Mcgraw-Hill.
6. Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the royal statistical society*, series b (methodological), 41, 2, 190-219.
7. Judge, G. G. & Griffiths, W. E. & Hill, R. C. & Lutkepohl, H. & Lee, T. (1985). *Introduction to theory and practice of econometrics*. New York. John Wiley and Sons.
8. Kennedy, P. (2003). *A guide to econometrics*. Cambridge, Massachusetts. The MIT press.
9. Kutner, M. & Christopher, N. & Neter, J. & Li, W. (2005). *Applied linear statistical models.* The McGraw Hill, international edition.
10. Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics,* 6, 2, 461-464.