Computational Statistics
https://doi.org/10.1007/s00180-022-01228-1

ORIGINAL PAPER

Model selection using PRESS statistic

Ida Marie Alcantara[1] · Joshua Naranjo[2] · Yanda Lang[3]

Received: 12 September 2021 / Accepted: 11 April 2022

Abstract

The most popularly used statistic R^2 has a fundamental weakness in model building: it favors adding more predictors to the model because R^2 can only increase. In effect, additional predictors start fitting noise to the data. Other measures used in selecting a regression model such as R^2_{adj}, AIC, SBC, and Mallow's C_p does not guarantee that the model selected will also make better prediction of future values. To avoid this, data scientists withhold a percentage of the data for validation purposes. The PRESS statistic does something similar by withholding each observation in calculating its own predicted value. In this paper, we investigated the behavior of R^2_{PRESS}, and how it performs compared to other criterion in model selection in the presence of unnecessary predictors. Using simulated data, we found R^2_{PRESS} has generally performed best in selecting the true model as the best model for prediction among the model selection measures considered.

Keywords PRESS statistic · Model selection · Prediction

Ida Marie Alcantara
alcanti@wwu.edu

Joshua Naranjo
joshua.naranjo@wmich.edu

Yanda Lang
yanda.lang@temple.edu

[1] Department of Mathematics, Western Washington University, 516 High Street, Bellingham 98225, USA

[2] Department of Statistics, Western Michigan University, 1903 W Michigan Ave, Kalamazoo, MI 49008, USA

[3] Department of Epidemiology and Biostatistics, Temple University, 1301 Cecil B. Moore Ave, Philadelphia, PA 19122, USA

Published online: 03 May 2022

1 Introduction

In simple regression analysis, the coefficient of determination represented as R^2, measures the proportion of variation in the response variable Y that is accounted for by regression on X. This is calculated as

$$R^2 = 1 - \frac{SSE}{SST} \tag{1}$$

where SSE represents the sum of the squared differences between the actual and predicted value of Y, that is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

and SST is the sum of squared differences between Y and the average of Y values:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

This most popularly used statistic has a fundamental weakness in model building when used in multiple regression analysis: it favors adding more predictors to the model because R^2 can only increase. In effect, these additional predictors start fitting noise in the data. Other measures used in selecting a regression model among candidate models include R^2_{adj}, Akaike Information Criterion (AIC), Schwarz Bayesian Information Criterion (SBC), and Mallow's C_p.

The adjusted R^2 criterion denoted as R^2_{adj} is a measure similar to R^2, but charges a penalty for the number of predictors included in the model. In its calculation, the ratio between SSE and SST in equation (1) is replaced by the ratio of MSE and MST, that is,

$$R^2_{adj} = 1 - \frac{SSE/n-p}{SST/n-1} = 1 - \frac{MSE}{MST} \tag{2}$$

where p represents the number of predictors in the model (including the intercept). In this criterion, adding a predictor in the model decreases SSE, but the denominator degrees of freedom $n-p$ also decreases which, as explained by Tamhane and Dunlop (2000), may result in an increase in MSE if the reduction in SSE caused by the additional predictors does not compensate for the loss in error degrees of freedom. Since MST is constant, an additional predictor in the model may result to a decrease in R^2_{adj}, contrary to R^2 which will only increase whenever a predictor is added.

Measures based on information criterion such as AIC and SBC combines the information about SSE with the number of predictors in the model and the sample size in the calculation. The formulas for these two measures are:

$$AIC = n\,ln(SSE) - n\,ln(n) + 2p \tag{3}$$

$$SBC = n\,ln(SSE) - n\,ln(n) + p\,ln(n) \tag{4}$$

As shown in Eqs. (3) and (4), AIC and SBC differ only in terms of the penalty for adding predictors in the model. AIC adds a $2p$ penalty, while SBC adds a $p\,ln(n)$ as penalty. In using either of these criterion in model selection, the model with the lowest value is selected as the best model.

Mallow's C_p statistic compares models with different subset of parameters relative to the full model, and calculated using the formula

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + 2(p+1) - n \tag{5}$$

where SSE_p is the SSE of the reduced model, while $\hat{\sigma}^2$ is the MSE of the full model. Murtaugh (1998) mentioned in his analysis that in using this criterion, "a stepwise procedure is used to add or delete predictors until a minimum value of C_p is obtained." In comparing models through this measure, the model that minimizes the C_p criterion is recommended as the final model.

Although useful information can be derived from these measures, it does not guarantee that the model selected using any of these criterion will not have variables that just add noise to the model, or make better prediction of future values. This is due to the fact that all the observations in the dataset were utilized in model building, hence what it provides as information is how well the model predicts the current observations.

To avoid this, data scientists withhold a percentage of the data for validation purposes. The PRESS statistic does something similar by withholding each observation in calculating its own predicted value. In this approach, the residual $e_{(i)}$ is calculated by taking the difference between the response variable Y_i, and its predicted value when the i^{th} observation is excluded from model building, $\hat{Y}_{(i)}$. That is,

$$e_{(i)} = Y_i - \hat{Y}_{(i)}$$

Taking the sum of squares of these residuals yields the Prediction Error Sum of Squares (PRESS) advocated by Allen (1971), defined as

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2$$

Since this process of calculating the PRESS statistic becomes tedious and requires a huge amount of run-time when the sample becomes large, the i^{th} PRESS residual can be expressed as a weighted term of the ordinary residual as used by Landram et al. (2011)

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}$$

where h_{ii} represents the leverage value which quantifies the influence that the observed response Y_i has on its predicted value $\hat{Y}_i$. Substituting this expression in place of SSE in equation (1), a measure similar to R^2 that is based on PRESS can be derived as:

$$R^2_{PRESS} = 1 - \frac{\sum_{i=1}^{n} \frac{e_i^2}{(1-h_{ii})^2}}{\sum_{i=1}^{n} (Y_i - \bar{Y}_{(i)})^2}$$

The denominator in this expression is adapted from the research of Mediavilla et al. (2008) which is a modification of the existing formula for *SST* used in calculating R^2. Among candidate models, the one with the highest value of R^2_{PRESS} is the best model for prediction. The question of interest is, will an R^2 based on PRESS fix R^2's inherent weakness of fitting noise?

Existing Research Several researches have been conducted comparing the performance of R^2_{PRESS} with other measures used in model selection. In his research, Murtaugh (1998) compared stepwise regression based on partial F-tests, Mallows' C_p, SBC, and regression trees and assess the ability of the methods to differentiate meaningful predictors from noise variables, and to further select the models that best predicts future observations. Mediavilla et al. (2008) compared R^2_{PRESS} with R^2_{adj}, AIC, and SBC both in estimating the parameters of the model, as well as the relative efficiency of R^2_{PRESS} compared to the other methods. Landram et al. (2011) focused on the limits and properties of R^2_{PRESS} as well as its differences with the other measures used in model selection. In this research, we introduced a true model where the data is generated from, and add unnecessary predictors to the model. Our goal is to compare the performance of R^2_{PRESS} with other existing measures as R^2, R^2_{adj}, AIC, SBC, and Mallows' C_p in selecting the true model as the best model for prediction, under different scenarios.

Outline The remainder of this article is organized as follows. Section 2 discusses in detail the behavior of R^2_{PRESS} when unnecessary variables are added in the model. A comparison of the performance of R^2_{PRESS} with other existing measures used in model selection in the small sample case are presented in Section 3. The summary of the comparison among the model selection measures when the sample size is large is presented in Section 4. Model selection using real data is presented on 5. Finally, Section 6 gives the conclusions and further work being done for the study.

2 Behavior of R^2_{PRESS}

To gain a better insight regarding the behavior of R^2_{PRESS} in multiple linear regression, a simulated data set was used for exploratory analysis. We begin with generating a dataset consisting of a response variable Y, with 3 predictors based on a specified value of R^2. The linear model for this data is given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Simulation was done using several values of the β coefficients to determine if this would have an effect on the results of the analysis. Since similar results were found after setting different values for the β coefficients, we arbitrarily fix the values for the rest of the analysis as follows: $\beta_0 = 0.2$, $\beta_1 = 0.3$, $\beta_2 = 0.5$, and $\beta_3 = 0.6$. Two values of R^2, 0.5 and 0.8, were used to determine if different R^2 values would yield a difference in the results. The X's were drawn independently from a standard normal distribution, while the values of ϵ were drawn from a normal distribution with $\mu = 0$ and σ of 0.7 for the dataset with target R^2 of 0.5, and 0.175 for the dataset with target R^2 of 0.8. Since R^2 is not utilized in model selection for multiple linear regression, we compared the performance of R^2_{PRESS} with R^2_{adj}. The simulation was done 10,000 times using a sample of size 80, and the comparison boxplots of the values of R^2_{adj} and R^2_{PRESS} are shown in Fig. 1.

Boxplots show the values of R^2_{PRESS} are lower compared to the values of R^2_{adj}. The average value of R^2_{adj} for the dataset with a target R^2 of 0.50 was calculated to be 0.5017 while R^2_{PRESS} was only 0.4814 on the average. For the datasets where the target R^2 is set to 0.80, the average value of R^2_{adj} was calculated to be 0.8023 while R^2_{PRESS} was 0.7943 on the average.

We next considered the effect of adding unnecessary predictors in the model. We added an increasing number of unnecessary predictors (t), generated as permutation of numbers from 1 to $n = 80$, to investigate the behavior of R^2_{adj} and R^2_{PRESS}. The average values from 10,000 simulations at each value of t, the number of unnecessary predictors added in the true model, are summarized in Table 1.

Table 1 shows that R^2_{adj} maintains its value at the target R^2 of the model, despite increasing value of t, while R^2_{PRESS} decreases as t increases. Furthermore, Table 1 shows R^2_{PRESS}, unlike the other two measures, can also take negative values in the presence of too many unnecessary predictors in the model.

Figure 2 shows the separation between the R^2_{adj} and R^2_{PRESS} with increasing value of t. The diagonal line in the plot depicts the line where R^2_{adj} and R^2_{PRESS} would be equal. The values denoted in red on the plots comparing the values of R^2_{adj} and R^2_{PRESS} with increasing value of t depicts the values from when the target $R^2 = 0.5$, while blue represent the values when the target $R^2 = 0.80$. The plots clearly shows as we increase the number of unnecessary predictors in the model, the dots move further away from the line where $R^2_{adj} = R^2_{PRESS}$, for both values of R^2 considered.

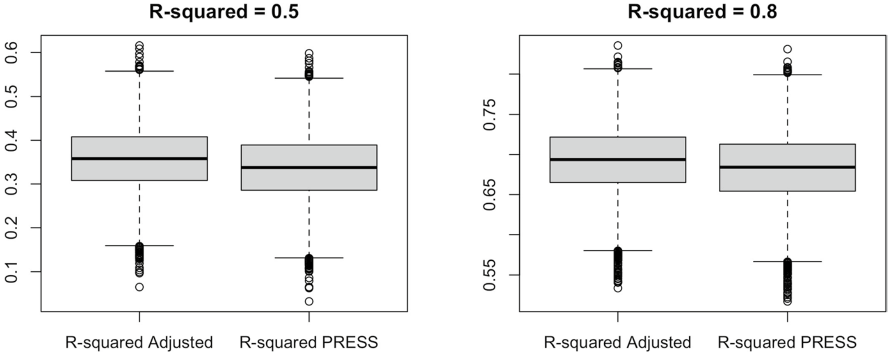
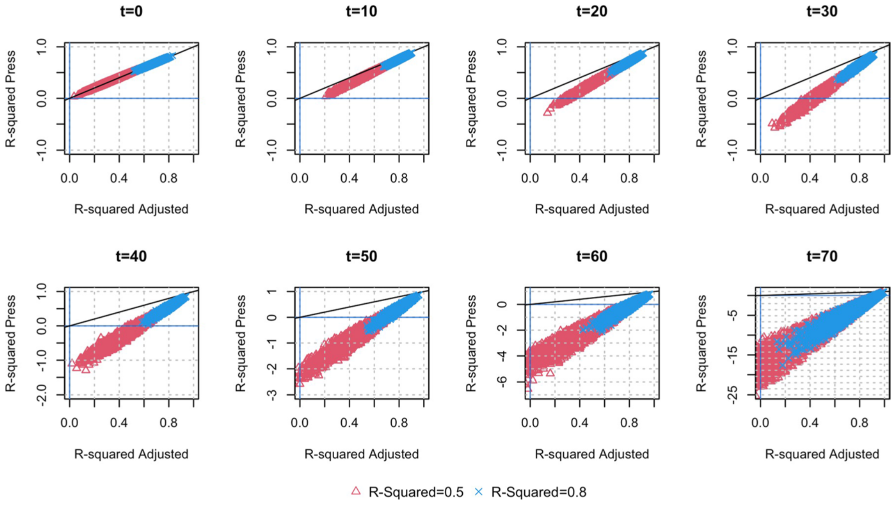


Fig. 1 Boxplots of R^2_{adj} and R^2_{PRESS} of the simulated data

Table 1 Average values of R^2_{adj} and R^2_{PRESS} against increasing t

	Target $R^2 = 0.50$		Target $R^2 = 0.80$	
t	R^2_{adj}	R^2_{PRESS}	R^2_{adj}	R^2_{PRESS}
0	0.5017	0.4814	0.8023	0.7943
1	0.5016	0.4743	0.8023	0.7915
2	0.5017	0.4671	0.8023	0.7886
3	0.5017	0.4599	0.8024	0.7858
4	0.5018	0.4523	0.8024	0.7827
5	0.5018	0.4445	0.8024	0.7796
10	0.5020	0.4024	0.8025	0.7630
20	0.5015	0.2935	0.8023	0.7198
30	0.5016	0.1353	0.8023	0.6570
40	0.5015	−0.1222	0.8023	0.5549
50	0.5022	−0.5836	0.8025	0.3719
60	0.5017	−1.7138	0.8024	−0.0764
70	0.5058	−8.117	0.8040	−2.6163


t=0

t=10

t=20

t=30

t=40

t=50

t=60

t=70

Fig. 2 Graphs of R^2_{adj} and R^2_{PRESS} against t

From these simulations, we can summarized the characteristics of R^2_{PRESS} as follows:

1. R^2_{PRESS} is always lower than R^2_{adj}.
2. Contrary to R^2_{adj} which maintains its value at the target R^2 with increasing number of unnecessary predictors, R^2_{PRESS} value decreases when the number of unnecessary predictors in the model increases.

3. R^2_{PRESS} can take on negative values in the presence of too many unnecessary predictors in the model.
4. The value of R^2_{PRESS} decreases faster when unnecessary predictors are added in the model, for datasets with smaller value of R^2.

These characteristics of R^2_{PRESS} shows the measure is more sensitive in identifying unnecessary predictors in the model compared to R^2_{adj} for the case where the predictors of the true model are uncorrelated.

3 Comparison of model selection measures: small sample

To get a better assessment of the performance of R^2_{PRESS} in model selection, we simulated datasets under different cases and compared the performance of existing optimality criteria such as R^2_{adj}, AIC, SBC, and Mallow's C_p, with that of R^2_{PRESS}. In this simulation study, we first considered the case when the sample size is small. We set the sample size to 15, and examined how well the measures in model selection will select the true model as the best model. Five unnecessary predictors were added to the true model, making the total number of predictors in the full model to be 8. Since the total number of predictors is relatively small, all subset regression was performed, and all $2^8 = 256$ were considered as candidate models.

In addition to the model selection criteria mentioned above, we also compared the results with that obtained by the *General-to-specific* (*GETS*) modeling. We used the *R* package *gets* by Pretis et al. (2018), which combines backwards elimination procedure with hypothesis tests on the β coefficients, diagnostic tests, as well as the information criterion in selecting the best fitted model among terminal models. The information criterion used in our simulation is the Schwarz information criterion, and the significance level α is set to 0.05.

We first examined the case where the X's of the true model and the errors are drawn from the Normal distribution. We considered the case where the X's are normal and uncorrelated, as well as when the X's are normal and correlated. The correlation of the X's was set to 0.25 for computational simplicity. Table 2 summarizes the percentage of correct identification made by each criterion out of the 10,000 simulations.

Table 2 shows R^2_{PRESS} and SBC have the highest percentage of correct identifications both for the correlated and uncorrelated X's case, while *GETS* modeling have the lowest percentage of correct identification. The percentages are also higher for all measures when the X's are uncorrelated, as compared to when the X's are correlated.

We next investigate the case when the X's drawn from the normal distribution are combined with errors drawn from a non-normal distribution. We considered a skewed distribution where $\epsilon \sim N + exp(\lambda = 0.25)$, and an extreme value distribution where $\epsilon \sim logexp$ as shown in Hettmansperger and McKean (2010). The results are summarized in Table 3.

Table 2 Percentage of correct identifications from 10,000 simulations ($n = 15$) for each criterion considered for the case where $X's \sim Normal$ and $\epsilon \sim Normal$

	Uncorrelated X's (%)	Correlated X's (%)
R^2_{adj}	6.34	4.25
R^2_{PRESS}	13.96	6.54
AIC	10.09	5.40
SBC	13.93	6.55
Mallow's C_p	12.20	6.26
GETS	2.31	1.82

When the errors are not normally distributed, the percentage of correct identifications made by each measure is much lower compared to the case when errors are normal. Although the percentages are roughly similar, R^2_{PRESS} has the highest percentage of correct identification among the model selection measures considered, with higher percentages observed when the errors are drawn from an extreme value distribution as compared to when the errors are from a skewed distribution.

The next case we looked into is that when the X's are drawn from a non-normal distribution. We considered drawing from a heavy-tailed distribution, as well as a skewed distribution for the true X's of the model in order to assess the performance of the measures in model selection. For the heavy-tailed distribution, the X's are drawn from a Logistic distribution, while for the skewed distribution, X's were drawn from a Skew Normal (SN) distribution with $\alpha = 5$. We combined these non-normal X's with errors drawn from Normal distribution and the Normal+Exponential distribution.

Table 4 shows higher percentages of correct identification for all the measures considered when $X's$ are drawn from heavy-tailed distribution, compared to when $X's$ are drawn from skewed distribution. For the Skew Normal case, R^2_{PRESS} still has the highest percentage of correct identification when the errors are normally distributed, but when errors are non-normal, R^2_{adj} and Mallow's C_p have higher percentages, although it is not that different from the percentage obtained by

Table 3 Percentage of correct identifications from 10,000 simulations ($n = 15$) for each criterion considered for the case where $X's \sim Normal$ and errors are non-normal

	$\epsilon \sim N + exp(\lambda = 0.25)$		$\epsilon \sim logexp$	
	Uncorrelated X's (%)	Correlated X's (%)	Uncorrelated X's (%)	Correlated X's (%)
R^2_{adj}	0.98	0.93	2.05	2.56
R^2_{PRESS}	1.26	0.95	2.75	4.15
AIC	1	0.76	2.15	3.08
SBC	0.88	0.70	2.16	3.14
Mallow's C_p	1.03	0.85	2.22	3.12
GETS	0.08	0.06	0.71	0.63

Table 4 Percentage of correct identifications from 10,000 (n = 15) simulations for each measure considered for the case where $X's$ are non-normal

	$X \sim SN$		$X \sim Logistic$	
	$\epsilon \sim N$ (%)	$\epsilon \sim N + exp(\lambda = 0.25)$ (%)	$\epsilon \sim N$ (%)	$\epsilon \sim N + exp(\lambda = 0.25)$ (%)
R^2_{adj}	2.30	0.059	9.16	2.48
R^2_{PRESS}	3.11	0.053	21.70	4.10
AIC	2.65	0.052	15.29	3.31
SBC	2.88	0.041	23.00	3.50
Mallow's C_p	2.74	0.059	15.24	3.30
GETS	1.05	0.031	11.02	0.08

R^2_{PRESS}. When X's are drawn from logistic distribution, R^2_{PRESS} have the highest percentage of correct identification when the errors are non-normal. In the case where errors are normally distributed, SBC has the highest percentage.

4 Comparison of model selection measures: large sample

We also examined the performance of model selection measures in the large sample setting with a large number of unnecessary predictors. The sample size was set to 80, and we kept the X's of the true model to 3, while adding up to 70 unnecessary predictors. Since all subset regression is computationally infeasible in this case, we considered only a number of candidate models to look at, and determined the percentage of correct identifications for each measure. Based on the all subset regression done for small samples, the smaller models have higher chance of selection as the best model hence we considered as candidate models all possible combination of X's of the true model, a small number of unnecessary predictors added in the model, then the increasing number of unnecessary predictors added in the model done in the earlier simulation. The list of candidate models considered in this section are summarized in Table 5.

Since *GETS* modeling will search through all 2^p possible models to find the best terminal model, we did not include the comparison with the *GETS* modeling in the large sample case. The summary of results for the case when the X's of the true model and the errors are drawn from the Normal distribution are summarized in Table 6.

When both the X's and errors are normally distributed, R^2_{PRESS} is second only to SBC in terms of the highest percentage of correct identifications, when X's are uncorrelated. However, when the X's are correlated, R^2_{PRESS} has the highest percentage among the measures considered.

We next looked at combining non-normal errors with Normal X's to examine how well the measures will identify the true model as the best model. Similar to the small sample case, the errors were drawn from a skewed distribution ($N + exp(0.25)$)

Table 5 List of candidate models considered for the large sample case

Model	Predictors in the model
1	X_1
2	X_2
3	X_3
4	X_1, X_2
5	X_1, X_3
6	X_2, X_3
7	X_1, X_2, X_3
8	X_1, X_2, X_3, t_1
9	X_1, X_2, X_3, t_1, t_2
10	$X_1, X_2, X_3, t_1 - t_3$
11	$X_1, X_2, X_3, t_1 - t_4$
12	$X_1, X_2, X_3, t_1 - t_5$
13	$X_1, X_2, X_3, t_1 - t_{10}$
14	$X_1, X_2, X_3, t_1 - t_{20}$
15	$X_1, X_2, X_3, t_1 - t_{30}$
16	$X_1, X_2, X_3, t_1 - t_{40}$
17	$X_1, X_2, X_3, t_1 - t_{50}$
18	$X_1, X_2, X_3, t_1 - t_{60}$
19	$X_1, X_2, X_3, t_1 - t_{70}$

Table 6 Percentage of correct identifications from 10,000 simulations ($n = 80$) for each measure considered for the case where $X \sim Normal$ and $\epsilon \sim Normal$

	Uncorrelated X's (%)	Correlated X's (%)
R^2_{adj}	5.81	5.38
R^2_{PRESS}	69.90	60.16
AIC	2.48	2.14
SBC	84.95	57.59
Mallow's C_p	47.76	39.91

and also from an extreme value distribution (*logexp*). The results are shown in the table below.

Table 7 shows that when X's are normally distributed but the errors are non-normal, R^2_{PRESS} have the highest percentage of correct identifications among the measures considered. The percentages are higher for uncorrelated X's than for correlated X's for all measures, and for the case when errors are skewed, Mallow's C_p performed better than SBC.

The last set of scenarios examined for the large sample case are those where the X's are drawn from a non-normal distribution. Similar to what was done for the small sample case, we investigated the performance of each measure when X's are drawn from Logistic distribution (heavy-tailed) and Skew Normal distribution (skewed), in combination with errors that are drawn from Normal

Table 7 Percentage of correct identifications from 10,000 simulations ($n = 15$) for each measure considered for the case where $X's \sim Normal$ and errors are non-normal

	$\epsilon \sim N + exp(0.25)$		$\epsilon \sim logexp$	
	Uncorrelated (%)	Correlated (%)	Uncorrelated (%)	Correlated (%)
R^2_{adj}	3.33	2.19	5.42	5.17
R^2_{PRESS}	22.24	13.71	50.09	46.62
AIC	0.79	0.54	1.86	1.81
SBC	7.58	2.68	39.87	33.21
Mallow's C_p	13.33	8.60	32.59	30.01

distribution and Normal+Exponential distribution. The summary of the results are shown in Table 8.

Table 8 shows among the cases considered, R^2_{PRESS} in general has the highest percentage of correct identifications compared to the other measures, except for the case when X's are heavy tailed and the errors are Normal, where SBC was able to correctly identify the true model as the best model for prediction 90.42% of the time.

5 Prediction of real data

To demonstrate how the model selection measures choose variables for prediction using real data, we used the highway accident rates data presented in the textbook of Weisberg (1985) and Hoffstedt's unpublished master's paper. The data consists of 13 predictors for accident rate taken from 39 sections of large highways in Minnesota in 1973 which are summarized in Table 9.

Analysis of the data found in the textbook by McQuarrie and Tsai (1998) shows that AIC, R^2_{adj}, and Mallow's C_p selected the best model with length of highway segment, speed limit, number of signalized interchanges per mile, number of access

Table 8 Percentage of correct identifications from 10,000 simulations ($n = 80$) for each measure considered for the case where X's are non-normal

	$X \sim SN$		$X \sim Logistic$	
	$\epsilon \sim N$ (%)	$\epsilon \sim N + exp(0.25)$ (%)	$\epsilon \sim N$ (%)	$\epsilon \sim N + exp(0.25)$ (%)
R^2_{adj}	5.83	1.38	6.72	5.10
R^2_{PRESS}	55.84	5.13	71.53	55.30
AIC	1.89	0.18	2.26	1.68
SBC	51.47	0.89	90.42	49.08
Mallow's C_p	37.01	3.34	49.99	35.81

Table 9 List of variables in the traffic accident data

Variable name	Description
ADT	Average Daily Traffic Counts (in thousands)
TRKS	Truck Volume as a Percent of the Total Volume
LANE	Total Number of Lanes of Traffic
ACPT	Number of Access Points per Mile
SIGS	Number of Signalized Interchanges per Mile
ITG	Number of Freeway-Type Interchanges per Mile
SLIM	Speed Limit in 1973
LEN	Length of the Highway Segment (in Miles)
LWID	Lane Width (in feet)
SHLD	Width of Outer Shoulder on the Roadway (in feet)
FA	Federal Aid Interstate Highway 1 if Yes, 0 otherwise
PA	Principal Arterial Highway 1 if Yes, 0 otherwise
MA	Major Arterial Highway 1 if Yes, 0 otherwise
RATE	1973 Accident Rate per Million Vehicle Miles

points, and principal arterial highway as predictors. In a conference paper presented by Ma (2017) which uses the same data, stepwise regression selected the best model with lane width, width of outer shoulder on the roadway, length of highway segment, and truck volume as predictors. We calculated the R^2_{PRESS} for this model and found that when these variables are included in the model, the value of R^2_{PRESS} is only 0.3126, which is even lower than the R^2_{PRESS} of the full model. We performed all subset regression on the data and identified the best model selected by each measure as well as the one obtained using *GETS* model, along with their corresponding R^2_{PRESS} values. The results of the analysis are summarized in Table 10.

Among the criterion considered, R^2_{adj}, AIC, and Mallow's C_p all selected the same model consisting of 5 predictors. The model selected by each measure have two common predictors: length of highway segment (LEN) and speed limit (SLIM). The number of signalized interchanges per mile (SIGS) were also identified as predictor by all measures considered except for SBC. Comparing the R^2_{PRESS} values of the model selected by each measure, difference in the R^2_{PRESS} values of the model selected by SBC and R^2_{PRESS} are not that different from each other, however, the variables included in the model are slightly different. The SBC model is a 3-predictor model which has an R^2_{PRESS} of 0.6236, while the model selected by R^2_{PRESS} has a slightly higher predictive ability with 5 predictors

Table 10 Summary of best model selected by each measure with their corresponding R^2_{PRESS} values

Measure	Variables in the model	R^2_{PRESS}
R^2_{adj}, AIC, Mallow's C_p	LEN SIGS SLIM ACPT PA	0.5722
R^2_{PRESS}	LEN TRKS SIGS SLIM PA	0.6466
SBC	LEN SLIM ACPT	0.6236
GETS	LEN SIGS SLIM LANE	0.5699

including truck volume (TRKS) which is consistent with the findings of other studies on traffic accidents such as the ones by Ma and Yuan (2018), and by Chang (2005). In both of these researches, truck volume appears as a contributor to traffic accidents, and among the measures examined in this study, only R^2_{PRESS} included truck volume as predictor in the final model, which demonstrates how competitive the measure is in model selection.

6 Conclusions

The research that we conducted provided a deeper understanding of the behavior of PRESS statistic and PRESS-based measures in multiple linear regression. For model selection, our simulation studies show the statistic R^2_{PRESS} is competitive with existing measures in model selection for multiple linear regression such as R^2_{adj}, AIC, SBC, and Mallow's C_p, as it was able to determine the true model as best model with higher percentage compared to other measures mentioned in most cases considered. In these simulation studies, R^2_{PRESS} has the highest percentage of correct identifications in cases where the X's are uncorrelated as well as when X's are drawn from skewed distribution. Our simulation studies also shows R^2_{PRESS} has better performance in determining the true model as best model when the errors are non-normal compared to the other model selection criteria. When the errors are normally distributed however, SBC has a better performance than R^2_{PRESS}. Comparing the models selected by each measure using a real data set, the model selected by R^2_{PRESS} includes variables that are consistent with existing models. Since R^2_{PRESS} is calculated through cross-validation, it provides better insight in evaluating the predictive ability of the model compared to the other measures considered in this study.

Our research is concentrated on how PRESS statistic performs compared to other model selection measures in linear regression setting. Further studies are being considered to determine how PRESS statistic can be extended to non-linear models, as well as how it performs compared to other cross-validation approaches. Comparison of the model selection measures including R^2_{PRESS} in the case when errors are autocorrelated is also under consideration, to get a full extent of the use of R^2_{PRESS} in model selection. Using Message Passing Interface (MPI) in R to include more candidate models in the large number of predictors case is also being considered.

References

Allen DM (1971) Mean square error of prediction as a criterion for selecting variables. Technometrics 13(3):469–475

Chang L-Y (2005) Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf Sci 43(8):541–557

Hettmansperger TP, McKean JW (2010) Robust nonparametric statistical methods, 2nd edn. CRC Press, Boca Raton, FL

Landram FG, Abdullat A, Shah V (2011) The coefficient of prediction for model specification. Southwest Econ Rev 32:149–156

Ma R (2017) The influence factors of highway traffic accident and accident rates model. In: Proceedings of 3rd international symposium on social science (ISSS 2017)

Ma W, Yuan Z (2018) Analysis and comparison of traffic accident regression prediction model. In: 3rd International conference on electromechanical control technology and transportation

McQuarrie AD, Tsai C-L (1998) Regression and time series model selection. World Scientific, Singapore

Mediavilla F, Landram F, Shah V (2008) A comparison of the coefficient of predictive power, the coefficient of determination and AIC for linear regression. J Appl Bus Econ 8(4):44

Murtaugh PA (1998) Methods of variable selection in regression modeling. Commun Stat Simul Comput 27(3):711–734

Pretis F, Reade JJ, Sucarrat G (2018) Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. J Stat Softw 86:1–44

Tamhane A, Dunlop D (2000) Statistics and data analysis: from elementary to intermediate. Prentice Hall, New Jersey

Weisberg S (1985) Applied linear regression. Wiley, New York

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.