Visit us - www.researchjournal.co.in ■ DOI : 10.15740/HAS/IRJAES/11.2/260-268

International Research Journal of Agricultural Economics and Statistics

Volume 11 | Issue 2 | September, 2020 | 260-268 ■ ISSN-2229-7278

Research Paper

Performance appraisal of validation techniques in R

■ **M. Iqbal Jeelani Bhat, Manish Kr. Sharma, Khalid-ul-Islam, Rizwan Yousuf and Zakir Hussain**

See end of the paper for authors' affiliations

Correspondence to :

Khalid-ul-Islam
Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu,
Jammu (J.&K.) India
Email: khalidstat34@gmail.com

Paper History :
Received : *16.05.2020;*
Revised : *25.07.2020;*
Accepted : *25.08.2020*

ABSTRACT : In this article various statistical models were fitted utilizing simulated symmetric and asymmetric data. Fitting of models were carried out with the help of various libraries like minpack.lm, matrices and nlme in R studio (version 3.5.1, 2018) and various selection criteria like RMSE, MAE, AIC, BIC were used for fitting of models. In order to evaluate different validation techniques the simulated data was divided in training and testing data sets and various functions in R were developed for the purpose of validation. Co-efficient summary revealed that all statistical models were statistically significant across both symmetric as well as asymmetric distributions. In preliminary analysis TFEM (Type First Exponential Model) was found out to be the best linear model across the distributions with lower values of RMSE, MAE, BIAS, AIC and BIC. Among non-linear models, Haung model was found out to be best model across both the distributions as it has lower values of RMSE, MAE etc. Different validation techniques like Half splitting, LOOCV and 5-folded cross validation were used in the present study. Based on the results of evaluation 5-folded cross validation performed better, as it resulted in lower rates of prediction error in comparison to its counter parts.

KEY WORDS : Model, Validation, R studio, Distribution, Prediction error rate

HOW TO CITE THIS PAPER: Bhat, M. Iqbal Jeelani, Sharma, Manish Kr., Khalid-ul-Islam, Yousuf, Rizwan and Hussain, Zakir (2020). Performance appraisal of validation techniques in R. *Internat. Res. J. Agric. Eco. & Stat.,* **11** (2) : 260-268, **DOI : 10.15740/HAS/IRJAES/11.2/260-268.** Copyright@2020:Hind Agri-Horticultural Society.

validation estimates of statistical models in terms of predictive performance are half splitting (50:50), leave one out cross validation (LOOCV) and k-folded cross validation.

Model validity is the stability and reasonableness of the regression co-efficients, Snee (1977) considered validation of regression models. For this he concluded that the data splitting is an efficient method of model validation when it is not practical to collect new data to test the model, one part of the data is used to estimate the model co-efficients and the rest of the data is used to measure the prediction accuracy of the model. Rencher and Run (1980) carried out set of Monte Carlo test to examine fake predictability and the inflation of R^2 as a function of sample size, size of the predictor and number of predictor selected. Anderson *et al.* (1982) made assessment and comparison of various functions by cross validation. Picard and Cook (1984) developed a methodology for assessment of the predictive ability of models and interest was given to models obtained via subset selection procedures and cross-validatory assessments of predictive ability were also obtained. Verbyla and Fisher (1989) developed a multiple regression site index model and reported that prediction bias potential due to over fitting a model with many predictor variables can be condensed by using cross-validation during model development. Model validation is an important step in the modelling process and helps in assessing the reliability of models before they can be used in decision making. Mayer (1992) stated that validation is an indispensable step for model approval. Efron (1993), examined the validation of models by comparing the predictive precision of data splitting techniques and bootstrapping approach to check the significance of each method in regression model validation and proposed a procedure for construction, selection and validation of regression models. Jeelani *et al.* (2015) evaluated linear regression model based on various sampling techniques like simple random sampling (SRS), systematic sampling (SYS) and rank set sampling (RSS) under LOOCV procedure. It was found that there was stability in the measure of R^2, Adj R^2 and RMSE in case of RSS as compared to SRS and SYS. Hassanzad *et al*. (2016) studied the relationship between height and diameter of velvet maple under cross validation and found 5-folded cross validation resulted in precise predication estimates. Jeelani *et al.* (2017) studied k-folded cross validation method for performance evaluation of different regression models. Jeelani *et al.* (2018) studied the relationship between fodder yield (dependent variable) and other parameters of *Grewia optiva* in Jammu region. In total, more than 30 models (including linear and non-linear) were evaluated and on the basis of adjusted R^2, the best five models were selected. Jeelani *et al.* (2018) studied the height diameter relationships of two parameter function models utilizing the data of 300 Pine trees and evaluated the models under 5-folded cross validation. In view of the above the present work was undertaken to gauge the performance of validation techniques in various types of statistical models and further, to suggest the competent validation technique in terms of predictive aptitude.

MATERIALS AND METHODS :

Both symmetric and asymmetric data was generated through simulation technique in R studio (R version 3.5.1) 2018 “Feather Spray”). R studio is an integrated development environment of the famous R software, which is a system for statistical analysis and graphics developed by Ross Ihaka and Robert Gentleman in the year 1995. R studio was developed on 28th of February 2011 by J.J Allaire who is an American software engineer (R Development Core Team, 2019). Following functions were developed to generate symmetric and asymmetric data sets.

Further the fitting of various linear and non-linear statistical models were carried out with the help of various libraries like library (minpack.lm, library (Metrics), library (caret, library (tidyverse) and library (nlme) available in R studio, also different functions were developed for fitting the models. The functional form and description of models used in the present study is given in table.

R codes developed for fitting of above mentioned statistical models are given below:

```
LM=lm(y1~x1,data=results1)
#function for first degree polynomial
PM= lm(log(y1)~log(x1),data=results1)
#function for power model
EM= lm(log(y1)~x1,data=results1)
#function for Type first exponential model
winsor1<- nlsLM(y1~a*(exp(-b)-exp(-c*x1)), data=results1,start=list(c=1,b=1,a=1))
#function for Winsor model
Grosenbaugh<- nlsLM(y1~a*(exp(-b/x1)+c),data= results1,start=list(a=1,b=1,c=1))
#function for Grosenbaugh model
```

Symmetric distribution	**Asymmetric distribution**
First degree polynomial	First degree polynomial
a<-1	a<-1
b <-1	b <-1
results1<- matrix(nrow=200,ncol=2)	results1<- matrix(nrow=200,ncol=2)
for (i in 1:200) {	for (i in 1:200) {
x1<- rnorm(200, mean=0, sd=1)	x1<- rexp(n=200, rate=.2)
e <- rnorm(200, mean=0, sd=1)	e <- rexp(n=200, rate=.3)
y1 <- a+b*x1+e	y1 <- a+B*x1+e
LM<- lm(y1~x1)	LM<- lm(y1~x1)
results[i,] <- coef(y1)	results[i,] <- coef(y)
}	}
LM	
Power model	Power model
a <- 1	a <- 1
b <-1	b <-1
results1 <- matrix(nrow=200,ncol=2)	results1 <- matrix(nrow=200,ncol=2)
for (i in 1:200) {	for (i in 1:200) {
x1<- rnorm(200, mean=0, sd=1)	x1<- rexp(n=200, rate=.2)
e <- rnorm(200, mean=0, sd=1)	e <- rexp(n=200, rate=.3)
y1=a*x1^b+e	y1=a*x1^b+e
PM<- lm(y1~x1)	PM<- lm(y1~x1)
results[i,] <- coef(y)	results[i,] <- coef(y)
}	}
Type first exponential model	Type first exponential model
a <- 1	a <- 1
B <-1	b <-1
results1 <- matrix(nrow=200,ncol=2)	results1 <- matrix(nrow=200,ncol=2)
for (i in 1:200) {	for (i in 1:200) {
x1<- rnorm(200, mean=0, sd=1)	x1<- rexp(n=200, rate=.2)
e <- rnorm(200, mean=0, sd=1)	e <- rexp(n=200, rate=.3)
y1=a*exp(x1*b)	y1=a*exp(x1*b)
EM<- lm(y1~x1)	EM<- lm(y1~x1)
results[i,] <- coef(y)	results[i,] <- coef(y)
}	}
Winsor Model	Winsor Model
a <- 1	a <- 1
b <-1	b <-1
c<-1	c<-1
results1 <- matrix(nrow=200,ncol=2)	results1 <- matrix(nrow=200,ncol=2)
for (i in 1:200) {	for (i in 1:200) {
x1<- rnorm(200, mean=0, sd=1)	x1<- rexp(n=200, rate=.2)
e <- rnorm(200, mean=0, sd=1)	e <- rexp(n=200, rate=.3)
y1= a*(exp(-b)-exp(-c*x1)),data=results1,start=list(c=1,b=1,a=1))	y1= a*(exp(-b)-exp(-c*x1)),data=results1,start=list(c=1,b=1,a=1))
winsor1 <-nls(y1~x1)	winsor1 <-nls(y1~x1)
results[i,] <- coef(y)	results[i,] <- coef(y)
}	}
Grosenbaugh Model	Grosenbaugh Model
a <- 1	a <- 1
b <-1	b <-1
c<-1	c<-1
results1 <- matrix(nrow=200,ncol=2)	results1 <- matrix(nrow=200,ncol=2)

Table : Contd………

Table : Contd……….	
for (i in 1:200) {	for (i in 1:200) {
x1<- rnorm(200, mean=0, sd=1)	x1<- rexp(n=200, rate=.2)
e <- rnorm(200, mean=0, sd=1)	e <- rexp(n=200, rate=.3)
y1= a*(exp(-b/x1)+c),data=results1,start=list(a=1,b=1,c=1))	y1= a*(exp(-b/x1)+c),data=results1,start=list(a=1,b=1,c=1))
Grosenbaugh<- nlsLM (y1~x1)	Grosenbaugh<- nlsLM (y1~x1)
results[i,] <- coef(y)	results[i,] <- coef(y)
}	}
Haung Model	Haung Model
a <- 1	a <- 1
b <-1	b <-1
c<-1	c<-1
results1 <- matrix(nrow=200,ncol=2)	results1 <- matrix(nrow=200,ncol=2)
for (i in 1:200) {	for (i in 1:200) {
x1<- rnorm(200, mean=0, sd=1)	x1<- rexp(n=200, rate=.2)
e <- rnorm(200, mean=0, sd=1)	e <- rexp(n=200, rate=.3)
y1= (a/(1+(b^-1)*(x1)^-c)),data=results1,start=list(a=1,b=1,c=1))	y1= (a/(1+(b^-1)*(x1)^-c)),data=results1,start=list(a=1,b=1,c=1))
Haung<- nlsLM (y1~x1)	Haung<- nlsLM (y1~x1)
results[i,] <- coef(y)	results[i,] <- coef(y)
}	}

Type	Name	Functional form	Description
Linear	FDM	Y= a + bX	Parameter a is the Y- intercept and it controls the vertical position of line. Parameter b is the slope of the line.
	PWRM	$Y= aX^b$	Also known as allometric function model. When parameter b is an integer, power function tends towards polynomial of degree b.
	TFEM	$Y = ae^{bX}$	Parameter a is the Y- intercept and parameter b is the shape of the curve. Widely used in forestry
Non – Linear	WNM	$Y = ae^{-be^{-cX}}$	It is mostly used by forest managers.
	GSRM	$Y= ae\ (-b/X^{+c})$	It is widely used for studying height diameter relationships.
	HAM	$Y=\left(\frac{a}{1+b^{-1}X^{-c}}\right)$	It includes number of models, also used widely used in forestry.

Haung <- nlsLM(y1~(a/(1+(b^-1)*(x1)^-c)),data= results1,start=list(a=1,b=1,c=1))

#function for Haung model

In order to get the summary of fitted models following in build functions of R were used :

#summary(name of model)

# where name of model can be (PM, EM, winsor1, Grosenbaugh, (Haung)

The adequacy of the fitted models were tested using different selection criteria like adjusted R^2 (R^2_{adj}), AIC, etc. Some the common metrics which were used in this stusy are as under:

$$R^2\ adj = 1-(1-R^2)\ x\left(\frac{n-1}{n\ x\ k-1}\right)$$

$$RMSE=\sqrt{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\big/_n}$$

$$MAE=\frac{\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|}{n}$$

AIC= n x In (RMSE) +2k

$$ME=\sum_{i=1}^{n}\left(\frac{yi-\hat{y}i}{n}\right)$$

Where, n is the number of observations, y_i is the actual observation, $\hat{y}_i$ the predicted value and $\bar{y}$ is the mean of observed value and k is number of parameters. R^2 Adj is adjusted co-efficient of determination, RMSE- Root mean square error, AIC- Akaike infor-mation criterion, MAE- Mean absolute error and ME- Mean error. Following functions of R studio utilizing library metrics were used to get the results for above mentioned selection criteria.

AIC (name of model)

BIC (name of model)

bias [y1, predict (name of model)]

mae [y1, predict (name of model)]

mape [y1, predict (name of model)]

rmse [y1,predict (name of model)]

# where name of model can be (LM,PM, EM, winsor1, Grosenbaugh, (Haung).

In order to assess the predictive performance of the models validation techniques like half splitting (50: 50), Leave one out cross validation (LOOCV) and 5-folded cross validation were used and various functions

Half splitting	**LOOCV and 5-folded cross validation**
mijbfjw.trainingsample<- results1$y1 %>%	
createDataPartition(p = 0.5, list = FALSE)	# Define training control
# results1 is the name of the simulated data frame (symmetric/asymmetric)	train.control<- trainControl(method = " ")
train.data<- results1[mijbfjw.trainingsample,]	# method = “LOOCV”/ “cv”
test.data<- results1[-mijbfjw.trainingsample,]	# Train the model
train.data	linearModel<- train(y1~., data = results1, method= "lm",
dim(train.data)	trControl = train.control)
test.data	powerModel<- train(log(y1)~log(x1), data = results1, method = "lm",
dim(test.data)	trControl = train.control)
linearModel<- lm(y1 ~x1, data = train.data)	EXPOModel<- train(log(y1)~ x1, data = results1, method = "lm",
summary(linearModel)	trControl = train.control)
powerModel<- lm(log(y1) ~log(x1), data = train.data)	WM Model <- train(nlsLM(y1~a*(exp(-b)-exp(-c*x1)),data = results1,
summary(powerModel)	method = "nlme",trControl = train.control)
EModel<- lm(log(y1) ~x1, data = train.data)	GM Model <- train(nlsLM(y1~a*(exp(-b/x1)+c),start=list(a=1,b=1,c=1),
summary(EModel)	data = results1, method = "nlme",trControl = train.control)
WM <- nlsLM(y1~a*(exp(-b)-exp(-c*x1)),data =	HM Model <- train(nlsLM(y1~(a/(1+(b^-1)*(x1)^-
train.data,start=list(c=1,b=1,a=1))	c)),start=list(a=1,b=1,c=1), data = results1, method = "nlme",trControl =
summary(WM)	train.control)
GM<- nlsLM(y1~a*(exp(-b/x1)+c),start=list(a=1,b=1,c=1), data = train.data)	# Summarize the results
summary(GM)	print(MODEL)
	# (where MODEL = linearModel, powerModel, EXPOModel, WM
Haung<- nlsLM(y1~(a/(1+(b^-1)*(x1)^-c)),start=list(a=1,b=1,c=1), data =	Model, GM Model & HM Model)
train.data)	
summary(Haung)	
# Make predictions and compute the RMSE, MAE, BIAS, AIc, BIC	
predictions<- MODEL %>% predict(test.data)	
# (where MODEL = linearModel, powerModel, EModel, WM, GM &Haung)	
data.frame(RMSE = RMSE(predictions, test.data$y1),	
MAE = MAE(predictions, test.data$y1),	
bias =bias(predictions, test.data$y1),	
AIC =AIC(predictions, test.data$y1),	
BIC=BIC (predictions, test.data$y1)	
PER <-RMSE(predictions, test.data$y1)/mean(test.data$y1)	

were developed for the purpose of their evaluation in R studio, which are given below:

RESULTS AND DATA ANALYSIS :

The summary statistics of the symmetric and asymmetric data generated through simulation are given in Table 1. The overall summary of the co-efficients of statistical models under symmetric and asymmetric distribution has been presented in Table 2 and 3. A perusal of these tables revealed that all the co-efficients of the statistical models were statistically significant which is an indication that models are well fitted across both the distributions. Table 4 and 5 revealed the performance of linear models across symmetric and asymmetric distribution utilizing various selection criteria like RMSE, MAE, BIAS, AIC and BIC. Among linear models type first exponential model (TFEM) was found out to be the best linear model across both symmetric and asymmetric distribution with lower values of RMSE, MAE, BIAS, AIC and BIC, which indicates that this model performs good in both symmetric as well as asymmetric data sets. Table 5 revealed the performance of non-linear models across symmetric and asymmetric distribution utilizing various selection criteria. Among non-linear models, Haung was found out to be the best non-linear model across both distributions as reflected from the values of RMSE, MAE etc which indicates that this model performs better in both symmetric as well as asymmetric distributions.

Different validation techniques like half splitting, LOOCV and 5-folded cross validation were used in this study. Table 6 reveal the performance criteria of FDM, PWRM and TFEM under various validation techniques

Table 1: Summary statistics of simulated data

	Variables	Mean	Median	Skewness	Kurtosis	Shapiro wilk test	
						W	p-value
Symmetric	Y1	0.9163	0.8932	0.1572	3.0197	0.9965	0.9395
Distribution	X1	0.0383	0.0314	0.0653	3.0386	0.9918	0.3295
Asymmetric	Y1	9.2395	8.1061	1.4103	5.7951	0.8983	1.986e-10
Distribution	X1	4.9134	3.5108	1.9795	8.9297	0.8136	9.96e-15

Table 2: Parameter estimates of linear models under symmetric and asymmetric distribution

Distribution	Model	Model equation	a	b
Symmetric	TFEM	$Y = ae^{bX}$	0.62*	0.10**
	PWRM	$Y= aX^{b}$	0.41	0.73**
	FDM	$Y=a+bX$	0.06*	0.06*
Asymmetric	TFEM	$Y= ae^{bX}$	0.02	1.31*
	PWRM	$Y= aX^{b}$	0.01	1.16**
	FDM	$Y= a+bX$	0.11	5.52**

(FDM = First degree polynomial model, PWRM= Power model, TFEM Type I Exponential Model)
*and ** indicate significance of values at P=0.05 and 0.01, respectively

Table 3: Parameter estimates of non linear models under symmetric and asymmetric distribution

Distribution	Model	Model equation	a	b	c
Symmetric	HAM	$\mathbf{Y = \left(\frac{a}{1+b^{-1}X^{-c}}\right)}$	3.14*	1.16**	0.72*
	WNM	$Y = ae^{-be^{-cX}}$	1.70**	1.35*	1.42**
	GSRM	$\mathbf{Y=ae\ (-b/X^{+c})}$	1.34**	0.53**	1.44*
Asymmetric	HAM	$\mathbf{Y = \left(\frac{a}{1+b^{-1}X^{-c}}\right)}$	5.14**	2.16*	0.72**
	WNM	$Y = ae^{-be^{-cX}}$	2.21*	0.84**	2.70*
	GSRM	$\mathbf{Y=ae\ (-b/X^{+c})}$	1.82*	0.16*	1.85**

(WNM = Winsor model, GSRM = Grosenbaugh model, HAM Haung model) * and ** indicate significance of values at P=0.05 and 0.01, respectively

Table 4: Performance criteria for linear models under symmetric and asymmetric distribution

Distribution	Model	Model equation	RMSE	MAE	BIAS	AIC	BIC
Symmetric	TFEM	$Y=ae^{bX}$	4.23	3.22	0.82	594.01	603.90
	PWRM	$Y= aX^{b}$	5.72	3.97	1.74	617.53	639.48
	FDM	$Y= a+bX$	6.96	3.99	1.95	1150.63	1161.58
Asymmetric	TFEM	$Y=ae^{bX}$	0.96	0.79	0.30	162.93	178.92
	PWRM	$Y=aX^{b}$	1.07	0.88	0.56	557.83	563.78
	FDM	$Y= a+bX$	1.13	0.92	0.74	605.12	625.27

(FDM = First degree polynomial model, PWRM= Power model, TFEM Type I Exponential Model)

Table 5 : Performance criteria for non-linear models under symmetric and asymmetric distribution

Distribution	Model	Model equation	RMSE	MAE	BIAS	AIC	BIC
Symmetric	HAM	$\mathbf{Y}=\left(\frac{\mathbf{a}}{1+\mathbf{b}^{-1}\mathbf{X}^{-\mathbf{c}}}\right)$	2.99	2.35	0.98	276.56	288.63
	WNM	$Y=ae^{-be^{-cX}}$	5.17	4.12	2.39	359.03	377.16
	GSRM	$\mathbf{Y=ae\ (-b/X^{+c})}$	5.23	4.31	2.63	384.11	397.05
Asymmetric	HAM	$\mathbf{Y}=\left(\frac{\mathbf{a}}{1+\mathbf{b}^{-1}\mathbf{X}^{-\mathbf{c}}}\right)$	4.18	3.17	0.00033	1141.67	1153.87
	WNM	$Y=ae^{-be^{-cX}}$	4.24	3.25	0.00045	1154.39	1168.55
	GSRM	$\mathbf{Y=ae\ (^{-b}/X^{+c})}$	4.27	3.28	0.00069	1179.90	1185.72

(WNM= Winsor model, GSRM= Grosenbough model, HAM Haung model)

Table 6 : Performance criteria of linear models utilizing different validation techniques under symmetric and asymmetric distributions

	Models	Validation	RMSE	MAE	BIAS	AIC	BIC	PER
Symmetric	FDM	50:50	0.90	0.54	0.26	144.91	163.52	1.49
		LOOCV	0.88	0.42	0.18	144.15	160.09	1.30
		5-FOLDED	0.61	0.33	0.08	102.95	116.33	0.55
	PWRM	50:50	0.99	0.72	0.29	435.85	464.03	1.11
		LOOCV	0.91	0.50	0.20	433.91	440.78	1.08
		5-FOLDED	0.76	0.39	0.14	360.58	400.75	0.66
	TFEM	50:50	1.03	0.79	0.094	468.69	477.92	0.82
		LOOCV	0.97	0.58	0.06	453.51	469.17	0.77
		5-FOLDED	0.79	0.41	0.03	365.11	379.54	0.42
Asymmetric	FDM	50:50	5.52	3.97	1.35	180.77	291.33	1.85
		LOOCV	3.91	3.08	1.02	195.72	221.23	1.51
		5-FOLDED	3.46	2.67	0.76	165.15	183.27	0.79
	PWRM	50:50	5.50	3.92	0.75	924.98	937.21	2.75
		LOOCV	3.86	2.81	0.26	670.89	685.49	2.67
		5-FOLDED	3.26	2.39	0.19	410.55	423.69	1.29
	TFEM	50:50	4.03	3.18	0.35	577.01	595.29	2.92
		LOOCV	3.52	2.73	0.22	560.58	583.69	2.91
		5-FOLDED	3.001	2.16	0.16	414.88	428.71	1.86

(FDM = First degree polynomial model, PWRM= Power model, TFEM Type I Exponential Model) in case of symmetric and asymmetric distribution. A perusal of Table 6 revealed that 5- folded cross validation performs better in comparison to half splitting and LOOCV across all the three linear models in case of symmetric distribution as it revealed lower prediction error rate (PER). Same results were found in case of asymmetric distribution.Table 7 revealed the performance criteria of HAM, WNM and GRSM under various validation techniques in case of symmetric and asymmetric distribution. A perusal of the above mentioned table again revealed that 5 folded cross validation performs better in comparison to its counter parts.

All the estimates of models used were significant, which means all the models were well fitted and total of 200 observations were simulated with respect to symmetric and asymmetric distributions. Among linear models, based on selection criteria Type first exponential model was found to be best linear model in both symmetric as well as asymmetric datasets as it has the lowest values of RMSE, MAE, BIAS, AIC and BIC. Amid non-linear models, based on selection criteria Haung model was found to be best non-linear model in both symmetric as well as asymmetric datasets because it has the lowest values of RMSE, MAE, BIAS, AIC and BIC. Under validation methods, in case symmetric distribution type first exponential model was found to be best linear model as it has the lowest prediction error rate (0.42), while as in asymmetric distribution first degree polynomial model was found to be best linear model as it has the lowest prediction error rate (0.79). On applying validation methods, in case symmetric distribution Winsor model was found to be best non-linear model as it has the lowest prediction error rate (0.87), while as in asymmetric distribution Haung model was found to be best linear model as it has the lowest prediction error rate (0.33). As far as evaluation of validation techniques are concerned 5-folded validation was found to be best in comparison to its counter parts as it has lower prediction error rates. The prediction error rate varied from 0.42 to 1.49 in case of linear models and the lowest prediction error rate was found in 5-folded cross validation across all the linear model under symmetric distribution. Prediction error rates varied from 0.87 to 1.98 in case of non-linear models and again lowest prediction error rates were found in case of 5-folded cross validation under symmetric distribution. In case of asymmetric distribution prediction error rate speckled from 0.79 to 2.92 in case of linear models and the lowest prediction error rate was found in 5-folded cross validation.Underneath asymmetric distribution prediction error rate mottled from 0.33 to 1.39

Table 7 : Performance criteria of non linear models utilizing different validation techniques under symmetric and asymmetric distributions

	Models	Validation	RMSE	MAE	BIAS	AIC	BIC	PER
Symmetric	HAM	50:50	4.09	3.64	2.91	228.59	373.49	1.27
		LOOCV	4.001	2.93	1.56	180.01	334.65	1.16
		5-FOLDED	3.84	2.64	0.77	140.91	299.33	0.96
	WNM	50:50	4.77	3.66	2.28	397.54	310.36	1.82
		LOOCV	4.52	3.33	1.71	365.27	297.11	1.51
		5-FOLDED	2.69	2.53	0.81	314.53	254.99	0.87
	GSRM	50:50	3.52	2.97	1.98	144.91	291.33	1.98
		LOOCV	2.91	2.08	1.637	144.15	221.23	1.75
		5-FOLDED	2.46	1.67	1.29	102.95	183.27	1.06
Asymmetric	HAM	50:50	2.50	1.78	0.09	196.51	217.89	0.96
		LOOCV	2.39	1.54	0.06	161.32	198.51	0.79
		5-FOLDED	1.77	1.01	0.001	102.27	117.29	0.33
	WNM	50:50	3.92	3.01	0.09	263.01	281.27	0.93
		LOOCV	3.66	2.59	0.05	260.11	273.09	0.63
		5-FOLDED	2.12	2.21	0.01	198.06	205.33	0.42
	GSRM	50:50	0.90	0.54	0.26	90.77	163.52	1.39
		LOOCV	0.89	0.42	0.18	67.72	140.09	1.12
		5-FOLDED	0.61	0.33	0.07	56.15	116.33	0.59

(WNM = Winsor model, GSRM = Grosenbaugh model, HAM Haung model) in case of non-linear models and again the results of prediction error rate were in favour of 5-folded cross validation in case of non-linear models.

Conclusion:

Hence, it is concluded 5- folded cross validation should be preferred whenever we have choice, because it gives lower prediction error rate, also it evaluates the model performance on different subsets, of the training data and then calculates the average prediction error rate. In contrast to LOOCV and jackknife, where model performance is tested at each iteration, which results in higher prediction error rates in former and higher values of BIAS in later, especially when data points are outliers, 5-folded cross validation provides solution under such circumstances by taking a good ratio of testing data points. Also the reason behind the lower rates of prediction error in 5-folded cross validation in comparison to half splitting is that every subset of data is used as training as well as testing data. This study can be a benchmark for policy makers, as formulation and initiation of economic policy and planning becomes easy if data sets are analyzed in advance which requires fitting and validation of various statistical models.

Authors' affiliations:
M. Iqbal Jeelani Bhat, Saquib Khan, Rizwan Yousuf and Zakir Hussain, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, **Jammu (J.&K.) India**

LITERATURE CITED :

Anderson, S., Madsen, S. F. and Rudemo, H. (1982). Examination and comparison of tree volume functions by cross-validation. *Forstlige Forsogsvaeseen Danmark*, **38**(3): 273-285.

Efron, B. (1993). Introduction to the Bootstrap. *Chapman and Hall.* London, United Kingdom.

Hassanzad, N. I., Alavi, S. J., Ahmadi, M. K. and Radkarmi, M. (2016). Comparison of different non-linear models for prediction of the relationship between diameter and height of velvet maple trees in natural forests (Case study: Asalem Forests, Iran). *J. Forest Science,* **62**(2): 65–71.

Jeelani, M. I., Mir, S. A., Khan, I., Nazir, N and Jeelani, F. (2015). Rank set sampling in improving the estimates of simple regression model. *Pakistan J. Statistics & Operation Research,* **11** (1): 39-49.

Jeelani, F., Sharma, M.K., Rizvi, S.E.H and Jeelani, M.I. (2017). Predictive modelling and validation for estimating fodder yield of *Grewia optiva*. *Malaysian J.Sci.*, **36**(2): 103-115.

Jeelani, M.I., Sharma, M.K., Bhat, A. and Gul, M. (2018). Validation of two parameter function height diameter models. *Internat. Res. J. Agric. Eco. & Stat.,* **9** (2): 331-334.

Jeelani, F., Sharma, M. K., Rizvi, S. E. H and Jeelani, M. I. (2018). A study on cross validation for model selection and Estimation. *Internat. J. Agric. Sci.*, **14**(1): 165-172.

Mayer, D.G. (1992). Statistical validation. *Ecological Modelling*, **68** (2) : 21-32.

Mosteller, F. and Turkey, J.W. (1968). *Data analysis, including statistics. In Handbook of Social Psycholog,* Addison-Wesley. pp. 601–720.

Picard. R.R. and Cook, R.D. (1984). Cross-validation of regression models. *J. American Statistical Association,* **79** (1) : 575-583.

R Development Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Rencher, A.C. and Run, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics,* **22** : 49- 53.

Snee, R.D. (1977). Validation of regression models: Methods and examples. *Technometrics*, **19**: 415-428.

Verbyla, L. D. and Fisher, F. R. (1989). An alternative approach to conventional soil–site regression modeling. *Canadian J.Forest Research*, **19** (2) : 179-184.