DOI: 10.15415/mjis.2013.12012

A Teaching Note for Model Selection and Validation

K. Muralidharan

Department of Statistics, Faculty of Science
The Maharajah Sayajirao University of Baroda,
Vadodara 390 002, India.
E-mail: lmv_murali@yahoo.com

Abstract

The model selection problem is always crucial for any decision making in statistical research and management. Among the choice of many competing models, how to decide the best is even more crucial for researchers. This small article is prepared as a teaching note for deciding an appropriate model for a real life data set. We briefly describe some of the existing methods of model selection. The best model from the two competing models is decided based on the comparison of limited expected value function (LEVF) or loss elimination ratio (LER). A data set is analyzed through MINITAB software.

1 INTRODUCTION

It is often believed that only one reasonable model may be constructed for a given decision making problem in market research. In empirical studies, one may be able to construct alternative models consistent with the hypothesis of the objectives (Wilson (1979), Bass (1969) and Carmone and Green (1981)). Many criteria may be used to compare quantitative marketing models. These criteria include such things as underlying assumptions, data requirements, and theoretical implications (Larreche and Montgomery (1977), Little (1979), Narasimhan and Sen (1983) and Rust (1981)).

Many methods for comparing structural forms of quantitative models have been proposed in the last three decades. These methods may be classified in terms of methodological emphasis as supermodel methods (Atkinson (1969), Johnson and Kotz (1977)), cross-validation methods (Mosteller and Tukey (1968), Stone (1974)), likelihood methods (Cox (1962), Akaike (1974)), or Bayesian methods (Smith (1973), Blattberg and Sen (1975), Rust (1981)). The model validation is a procedure of model comparison, estimating the competing models and then compiling error statistics on the other (Mosteller and Tukey (1968), Stone (1974)). The interrelatedness of the above categories of methods is exemplified by the fact that using Bayesian arguments Smith and Spiegelhalter (1980) obtained criteria closely related to those of Akaike (1974) and Schwarz information criterions.

This article examines the case in which two or more probability models are to be compared and decide how to choose the best among the competing

Mathematical Journal of Interdisciplinary Sciences
Vol. 1, No. 2,
March 2013
pp. 55–62

models. In the next section, we present the problem. In section three some useful results and model inferences are discussed.

2 THE PROBLEM

Consider the following data set (n = 40): 85, 90, 92, 5, 10, 17, 54, 55, 58, 55, 58, 32, 33, 32, 82, 68, 34, 36, 92, 102, 103, 106, 146, 124, 142, 195, 65, 66, 68, 54, 55, 58, 143, 151, 158, 195, 114, 114, 116 and 57 (see Bain and Engelhardt, 1991). Assume that this is not a time series data, then as a preliminary investigation, one can prepare a frequency distribution or a histogram or even a sample ogive to get an assessment about the shape of the data. The ogive and histogram generally provide some indication as to the type of distribution that will model the data. For the above data set we prepared a histogram and ogive (omitted here) and concluded to be a positive skewed distribution (the MINITAB sample statistics for the above data are mean = 83, median = 68, mode = 55, sd = 47.58 and skeweness = 0.614). Instead of drawing conclusions from descriptive statistics, one can directly check the normality of the data and proceed. On our further investigation, we found that the prospective models for the above data set are Weibull and Gamma distributions (see Table 2 for p-values). In the next section, we present some useful results to decide best alternative model through a data set.

3 SOME USEFUL RESULTS

Suppose X be the decision variable under study and $f(x)$ the corresponding probability density function, then the limited expected value (LEV) function or the expected loss eliminated is defined as

$$E[X;d] = \int_0^d xf(x)dx + d[1 - F(d)], \tag{1}$$

where $F(x)$ is the distribution function of the variable X. Another view of E[X;d] is that it is the expected value of $Y = \min(X, d)$, that is the mean of a random variable censored at d. The other quantity of interest is the loss elimination ratio (LER) which is the ratio of the expected loss eliminated to the expected value of X. that is LER = E[X;d] /E(X), provided E(X) exists (Klugman et.al. 2008). Note that E[X;d] always exists. A quantity which needs to be compared with E[X;d] is the empirical limited expected value (ELEV) function for a sample as

$$E_n(d) = \frac{1}{n}\sum_{i=1}^{n} \min(x_i, d) \tag{2}$$

For accepting any model as providing a reasonable description of the decision process, we should verify that E[$X;d$] and $E_n(d)$ are essentially in agreement for all values of d. It is because as $d \to \infty$ E[$X;d$] $\to$ E(X), if it exists and $E_n(d) \to \bar{X}$. Thus comparing E[$X;d$] and $E_n(d)$ is like a method of moments approach in a restrictive way.

In life testing experiments d may be sometimes called the truncation time or the censoring time. In this respect another important characteristics of life time models is the mean residual life (MRL) at age $d > 0$ is the conditional mean of X-d, given X $\geq d$, namely

$$e(d) = E[X - d \mid X \geq d] = \int_d^\infty (x-d)\frac{f(x)}{P(X \geq d)}dx \tag{3}$$

Then E[X;d] and e(d) are related through the equality

$$\text{E(X)=E[X;}d\text{]} + \text{e}(d)[1 - \text{F}(d)] \tag{4}$$

A plot of e(d) can also give some indication as to the type of distribution that will model the data. A very important problem associated with model selection is fitting of probable model to the data. In order to fit a model to a data, we need to estimate the parameters. This can be done using various methods like percentile matching, method of moments, minimum distance method, minimum chi-square method and maximum likelihood method etc. The first two are crude method and may be easy to implement but produce inferior estimates. The other three methods are more formal procedures with well defined statistical properties. Although they produce reliable estimators, they can be complex sometimes. Many often we need to employ some numerical procedures to get the estimators. For the particular problem discussed above, we use maximum likelihood method to estimate the parameters. In Table 1 we provide the MINITAB output of the model parameters estimates and Table 2 the corresponding Goodness of fit summary for the above data set.

In Table 2, we present the summary of Goodness of fit tests and their conclusion. The Table clearly shows a preference for Weibull distribution followed by Gamma distribution. It is possible to judge the best based on the p-value concept also. A p-value is a measure of how much evidence we have against the null hypotheses. It also measures the consistency by calculating the probability of observing the results from the sample of data assuming the null hypothesis is true. In the above situation, the p-value concept may not be sufficient to arrive at a final conclusion as the same models have same p-value.

Table 1. ML Estimates of Distribution Parameters

Distribution	Location	Shape	Scale
Normal*	83.00000		47.58636
Lognormal*	4.20502		0.76126
3-Parameter Lognormal	4.98178	0.30774	- 69.73051
Exponential			83.00000
2-Parameter Exponential	78.05000		4.95000
Weibull		1.80328	93.12422
Smallest Extreme Value	107.70310		50.63953
Gamma		2.49243	33.30081
3-Parameter Gamma	4.7124	22.26230	- 20.99262
Logistic	79.53017		27.02753
Loglogistic	4.28257		0.39303
3-Parameter Loglogistic	4.81433	0.21261	- 47.91801

Table 2. Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	0.633	0.092	
Lognormal	1.050	0.008	
3-Parameter Lognormal	0.313	*	0.002
Exponential	3.186	< 0.003	
2-Parameter Exponential	2.812	< 0.010	0.027
Weibull	0.317	> 0.250	
Smallest Extreme Value	1.486	< 0.010	
Largest Extreme Value	0.313	> 0.250	
Gamma	0.432	> 0.250	
3-Parameter Gamma	0.303	*	0.254
Logistic	0.593	0.083	
Loglogistic	0.589	0.085	
3-Parameter Loglogistic	0.352	*	0.056

Therefore, to decide between the two we now use a semi parametric approach based on the agreement between the empirical and fitted LEV function as given in equations (1) and (2) above. These functions are crucial to the determination of the effects of coverage modifications.

Figure 1

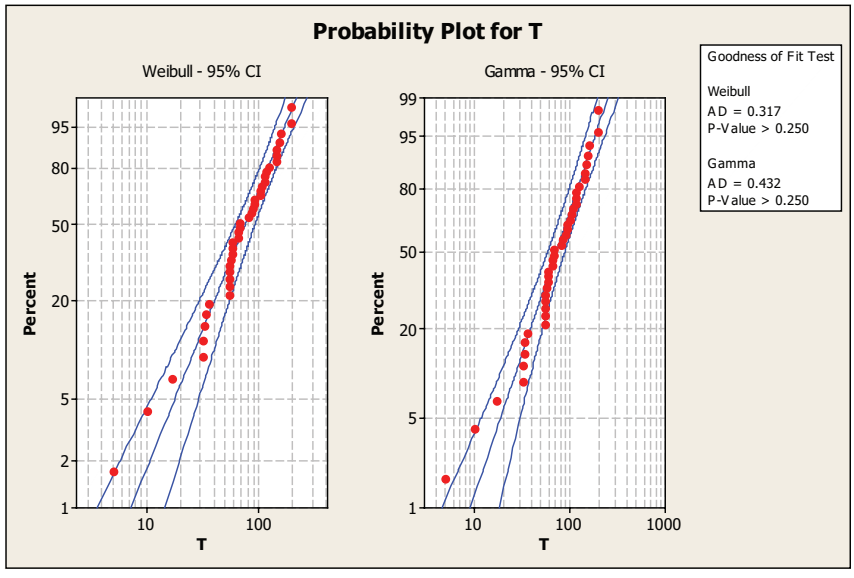

Table 3.

Empirical estimates.

X	$E_n(x)$	$F_n(x)$	Gamma		Weibull	
			E(Y;x)	$F_Y(x)$	E(Y;x)	$F_Y(x)$
5	5	1/40	4.85	.059	4.78	.226
10	7.5	2/40	9.42	.114	8.45	.617
20	10.67	3/40	17.76	.215	9.98	.976
40	24.87	8/40	31.72	.383	25.8	.999
80	40.10	20/40	51.27	.619	43.56	1.00
120	64.30	32/40	63.34	.766	62.44	1.00
150	72.14	37/40	69.22	.836	71.19	1.00
200	79.32	1	75.34	1.00	77.11	1.00

For the example discussed above, except Weibull and Gamma we rejected the other models because of their poor performance on likelihood measures. As Weibull and Gamma are the two competitors (see Figure 1), we computed the LEV functions for both distributions to make the final decision. In Table 3, we present those empirical estimates. From the table it is seen that for Weibull distribution the empirical expectation is closer to the LEV and hence is a better model for the data. Also the LER of Weibull is much less than that of the Gamma distribution for every d as the estimated E(X) of Weibull and Gamma is respectively 352.87 and 82.71. This again supports Weibull being the best model for the data set.

The table also includes the empirical distribution function which may be used for manually checking goodness-of-fit test. The above method works even if there are more than two competing models. The use of MRL function can also give a similar result. Since MRL and LEV functions are related it is sufficient to do only LEV function calculations. The other methods like minimum distance estimation, Bayes factor approach etc can also be used for identifying the best out of many alternatives. Generally these methods are very tedious.

ACKNOWLEDGEMENTS

The author thanks the Editor for valuable comments. The inspiration behind the preparation of this article is due to Prof. A. K. Chaudhari of ISI Bangalore on a model identification lecture (Six sigma MBB08).

REFERENCES

Atkinson, A. C. (1969). A method for discriminating between models, *Journal of the Royal Statistical Society* (B), 32, 323-353.

Akaike, H. (1974). A new look at the Statistical identification model. *IEEE Trans. Auto. Control,* 19, 716-723. http://dx.doi.org/10.1109/TAC.1974.1100705

Bain, L. J. and Engelhardt, M. E. (1991). Statistical analysis of reliability and life-testing analysis. Marcel Dekker, New York.

Bass, F. (1969). A simultaneous equation regression study of advertising and sales of cigarettes, *Journal of Marketing Research*, 6, 291-300.

Blattberg, R. C., and Sen, S. K. (1975). An evaluation of the application of minimum Chi-square procedures to stochastic models of brand choice, *Journal of Marketing Research,* 10, 421-427. http://dx.doi.org/10.2307/3149390

Carmone, F. H. and Green, P. E. (1981). Model misspecification in multi-attribute parameter estimation, *Journal of Marketing Research,* 18, 87-93.

Cox, D. R. (1962). Further results on tests of separate families of hypothesis, *Journal of the Royal Statistical Society* (B), 24, 406-424

Johnson, N. L. and Kotz, S. (1977). *Urn models and their applications*. John Wiley & Sons, New York.

Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2008). Loss Models: From Data to Decisions, (Third Edition). Wiley Series in Probability and Statistics.

Larreche, J. C. and Montgomery, D. B. (1977). Framework for the comparison of marketing models: A Delphi study. *Journal of Marketing Research*, November, 487-498.

Little, J. D. (1979). Models and Managers: The concept of a decision calculus. *Management science,* 16, B466-B485.

Mosteller, F. and Tukey, J. W. (1968). Data analysis, including statistics, in handbook of Social Psychology, F. Lindzey and E. Aronson (Eds.), Vol.2, Reading, Mass.: Addison-Wesley.

Narasimhan, C. and Sen, S. K. (1983). New product models of test market data. *Journal of Marketing,* 11-24.

Rust, R. T. (1981). A probabilitistic measure of model superiority. Working paper 81-23, Graduate school of Business, University of Texas, Austin.

Smith, A. F. M. (1973). A general Bayesian linear model, *Journal of the Royal Statistical Society* (B), 2, 213-220.

Stone, M. (1974). Cross-valedictory choice and assessment of Statistical predictions. *Journal of the Royal Statistical Society* (B), 36, 111-147.

K. Muralidharan is currently working as Professor and Head of department of Statistics, Faculty of Science, The Maharajah Sayajirao University of Baroda, Vadodara.