THE UNIVERSITY OF RHODE ISLAND

University of Rhode Island
DigitalCommons@URI

Psychology Faculty Publications

Psychology

1986

Goodness-of-fit patterns in a computer cross-validation procedure comparing a linear and a threshold model

Charles E. Collyer
University of Rhode Island, collyer@uri.edu

Follow this and additional works at: https://digitalcommons.uri.edu/psy_facpubs

Terms of Use
All rights reserved under copyright.

Citation/Publisher Attribution
Collyer, C. E. Behavior Research Methods, Instruments, & Computers (1986) 18: 618. https://doi.org/10.3758/BF03201437
Available at: https://doi.org/10.3758/BF03201437

This Article is brought to you for free and open access by the Psychology at DigitalCommons@URI. It has been accepted for inclusion in Psychology Faculty Publications by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons@etal.uri.edu.

Behavior Research Methods, Instruments, & Computers
1986, 18 (6), 618-622

STATISTICAL TECHNIQUES

Goodness-of-fit patterns in a computer cross-validation procedure comparing a linear and a threshold model

CHARLES E. COLLYER
University of Rhode Island, Kingston, Rhode Island

Cross-validation is the process of comparing a model's predictions to data that were not used in the estimation of model parameters. Cross-validation may have some value in identifying source models, especially in cases where the corresponding fitted models require the estimation of different numbers of parameters. Some of the information available from cross-validation is illustrated using a linear and a threshold model, and goodness-of-fit patterns are contrasted with those of conventional model-fitting.

Cross-validation can best be defined by contrasting it with conventional model-fitting. In the conventional procedure, a model equation is fit to data by using the data to estimate values for the model's free parameters and measuring the degree of resemblance between the model's predictions, given those parameters, and the same set of data. In cross-validation, there are two sets of data, independently sampled from the same source: the *estimator set* is the data from which model parameters are estimated, and the *criterion set* is the data with which model predictions are compared.

Cross-validation of simple-regression and multiple-regression models is considered a strong test of the predictive validity of specific regression equations. Confidence in an equation is enhanced if it can be shown to fit well in cross-validation, because although conventional goodness-of-fit may be inflated by capitalization on chance, cross-validation fits do not, in general, benefit from this random factor. One expects lower goodness-of-fit in cross-validation because the estimator and criterion data sets have mutually independent error components. Although lower goodness-of-fit is not inevitable, cross-validation is sometimes used as an empirical way to estimate shrinkage of the goodness-of-fit measure.

Although cross-validation has had some strong advocates in psychometrics (most notably Mosier, 1951), its use remains limited because the need to collect and analyze two sets of data places significant additional demands on a researcher's time and resources. The opinion is also expressed by some workers that it is better to use all available data in the estimation of model parameters than to risk unreliable estimates by using only half of the available data. There is merit in this view: in effect, cross-validation trades reliability for information about predictive validity, and this is a trade that must be weighed by those who are considering the costs and benefits of the procedure.

In experimental psychology, including the areas in which mathematical modeling is common, cross-validation is simply not in the repertoire of commonly used methods. Psychologists formulate models and fit them to data, as in conventional simple regression. A model's goodness-of-fit is expressed as r^2, if the model asserts linearity between the independent and dependent variables, or as an analogous proportion of variance accounted for, if the model is nonlinear. Goodness-of-fit may also be measured by a chi-square statistic. When models are compared, the best fitting model is taken to be a better characterization of the process underlying the data than those models that fit less well.

A consideration of how models capitalize on chance in conventional model-fitting, however, tends to complicate the interpretation of goodness-of-fit. In general, model predictions conform ever more closely to the data as more parameters are estimated from the data to be fit. A model with more parameters may fit a given set of data better than another model, not because it is more valid, but because it has a greater capacity to capitalize on the particular pattern of chance fluctuations in that data. This differential capitalization becomes relevant in research when the alternative hypotheses to be tested are captured in models that have different numbers of free parameters.

If models with more parameters can be expected to have an unfair advantage in conventional model-fitting, what can be said about cross-validation? The attraction of cross-validation is that no parameters are estimated from the

Portions of this work were presented at the meeting of the Psychonomic Society, Boston, November, 1985. I am indebted to L. J. Kamin, who suggested computing the correlations shown in Figure 3.

Address correspondence to Charles E. Collyer, Department of Psychology, University of Rhode Island, Kingston, RI 02881.

Copyright 1986 Psychonomic Society, Inc.

criterion data, so that conventional capitalization on chance does not differentially inflate goodness-of-fit. However, at least two potential disadvantages weigh against this attraction. First, cross-validation involves more work than does conventional model-fitting and may not be practical unless time, resources, and subjects are abundant. Second, models having different numbers of parameters are likely to display different degrees of shrinkage. For example, the model with more parameters may be expected both to conform more closely to the estimator data and to transfer less well to the criterion data, whenever the level of error in the data sets is such that parameter values obtained from the estimator set are influenced by capitalization.

Collyer (1985) recently studied the sensitivity of three fitted models to source-model variation in simulations of conventional model-fitting. The three models were constructed originally as hypotheses about the shape of the mental rotation response-time function. Two of these models were used in the present study of cross-validation procedures.

The linear model is a two-parameter model predicting a simple linear relation between response time and stimulus angle. The model equation is written

$$RT = sA+i+e, \qquad 0 \leq A \leq 180,$$

where RT is a single response time, A is the stimulus angle (in degrees), s is the slope of the predicted line, i is the intercept, and e is a random error component with an expected value of zero.

The threshold model is a four-parameter model predicting a nonlinear, two-part function:

$$RT = k+e, \qquad 0 \leq A \leq t,$$

$$RT = sA+i+e, \qquad t< A \leq 180,$$

where the new terms are t, the threshold for mental rotation, and k, the subthreshold response time. The threshold model expresses one version of the hypothesis that small (subthreshold) angles do not require mental rotation; thus, subthreshold response times are predicted to be independent of stimulus angle.

The specific purposes of this study can now be stated in relation to these models. First, when these two models are fit in a conventional way to mental rotation data, the threshold model must fit at least as well as the linear model regardless of its validity, because the linear model is a special case of the threshold model. This relationship between the two models invalidates the decision rule "Choose the best fitting model," at least in conventional model-fitting (Collyer, 1985). The computer simulations in the present study allowed this decision rule to be studied in cross-validation. Second, the degree to which a more complex model fits better in a conventional analysis often determines whether it will be accepted. This margin of superiority, however, tends to be inflated by capitalization on chance (Collyer, 1985). It seems possible, therefore, that the degree of superiority in conventional model-fitting is inversely related to predictive ability in cross-validation. One of the purposes of this study was to examine this possibility. Third, with information available about both conventional model-fitting and cross-validation for the same models, the study was an opportunity to provide an overview of these procedures and to summarize some principles of model identification.

METHOD

Specification of Source Models

Fixed parameter values for each model were chosen by averaging the estimates from 20 human subjects, who took part in an earlier mental rotation experiment (Rossi & Collyer, 1986). The linear source model was

$$RT = 14.4A+1034+e, \qquad 0 \leq A \leq 180. \quad (1)$$

The threshold source model with fixed parameter values was

$$RT = 1114+e, \qquad 0 \leq A \leq 21,$$

$$RT = 11.0A+1499+e, \qquad 21 < A \leq 180. \quad (2)$$

Equations 1 and 2 represent approximations to a typical subject's performance, under each of the models. The general goal of both the conventional and the cross-validation model-fitting operations was to detect empirical differences between data sets generated by Equations 1 and 2.

Five noise-level conditions were defined by setting the standard deviation of e equal to 42.5, 200, 425, 650, or 850. The middle value, 425, approximates the noise level in one session's worth of data from a human subject, as measured by the linear model's standard error of estimate.

Apparatus

An Apple II+ computer with 48K memory was used.

Computer Programs

Two simulation programs, titled XVAL LINEAR SOURCE and XVAL THRESHOLD SOURCE, were used to generate data and perform selected computations. The only difference between the two programs was in the source model underlying the simulated data. For convenience, data were collected and summarized in blocks of 100 simulated mental rotation subjects. Ten blocks were run at each of the 10 combinations of source model and noise level. The generation and analysis of data sets were as follows. The Box Muller algorithm (Box & Muller, 1958) was used to generate 16 random normal variates for angles of 0°, 3°, 6°, 9°, 12°, 15°, 18°, 21°, 24°, 27°, 30°, 60°, 90°, 120°, 150°, and 180°, with means given by the source model for each angle and standard deviation by the noise level. Two such data sets, the estimator and the criterion, were obtained for each subject. The product-moment correlation of the estimator and criterion

sets was calculated. The linear model was fit to the estimator set by simple regression to get the parameter estimates s, the slope, and i, the intercept, of the function. The threshold model was fit to the estimator data by finding the combination of values for the parameters k, t, s, and i that minimized residual variation (Collyer, 1985). Predictions of each fitted model were then compared to the estimator data for conventional model-fitting and to the criterion data for cross-validation.

RESULTS AND DISCUSSION

The mean correlation between the estimator and criterion data sets declined from about 1.0 at the lowest noise level to about 0.48 at the highest. At any noise level, this correlation was the same for both source models to the second or third decimal place; thus, there was no interaction between noise level and source model for this measure of estimator-criterion comparability.

A conventional goodness-of-fit summary of the estimator data is shown in Table 1. The fitted threshold model showed uniformly higher proportions of variance accounted for, because of its larger number of free parameters. Even the small margins of superiority in this table are significant, on rational as well as statistical grounds. Goodness-of-fit declined as a function of noise level for all source model/fitted model combinations.

A cross-validation goodness-of-fit summary of the criterion data is shown in Table 2. When the source model was linear, the fitted linear model performed better than the threshold fitted model, and the margin of superiority increased as a function of noise level. When the source of the data was the threshold model, the best fitting model

Table 1
Conventional Goodness-of-Fit: Proportions of Estimator Variance Accounted for by Model Predictions

	Noise Level				
Fitted Model	Very Low	Low	Medium	High	Very High
	Source: Linear Model				
Linear	.997	.94	.80	.63	.48
Threshold	.998	.96	.85	.71	.60
	Source: Threshold Model				
Linear	.94	.89	.74	.59	.49
Threshold	.998	.95	.83	.70	.63

Table 2
Cross-Validation: Proportions of Criterion Variance Accounted for by Model Predictions

	Noise Level				
Fitted Model	Very Low	Low	Medium	High	Very High
	Source: Linear Model				
Linear	.997	.94	.74	.53	.37
Threshold	.996	.92	.69	.45	.23
	Source: Threshold Model				
Linear	.93	.88	.67	.48	.28
Threshold	.997	.92	.67	.43	.17

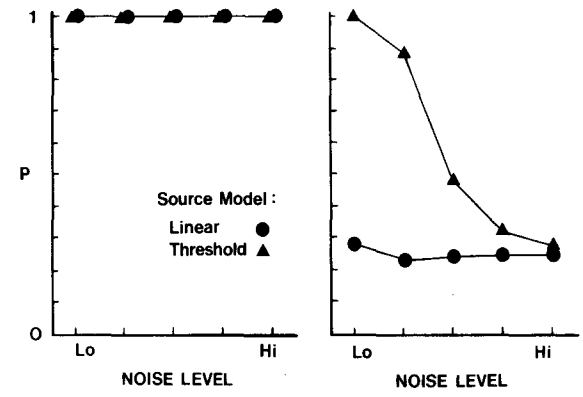

Figure 1. Probability that the threshold fitted model will fit better than the linear fitted model, as a function of source model and noise level. Left panel: conventional model-fitting. Right panel: cross-validation.

depended on the noise level; for the two lowest noise levels, the threshold fitted model was best, whereas for the two highest noise levels, the linear fitted model was best.

Which of the two fitted models gave the best fit under various conditions? In conventional fitting (to the estimator data sets), the threshold model always fits best, as shown in the left panel of Figure 1. In cross-validation, the relative frequency with which the threshold fit exceeds the linear fit, is a joint function of source model and noise level, as shown in the right panel of Figure 1. The probability that the threshold model will fit best when it is "true" (i.e., when the source model is Equation 2), is a rapidly declining function of noise level. The probability that the threshold model will fit best when it is "false" (i.e., when the source model is Equation 1), is virtually independent of noise level. These two curves are analogous to hit and false-alarm functions; the difference between the curves can be viewed as a measure of the discriminability of the threshold and linear source processes in cross-validation, using only best fit information.

The data of Figure 1 can be approached from a Bayesian perspective, with a view to finding the probability of correctly identifying the source model using the decision rule "Choose the best fitting model." In a conventional analysis, the probability that the source model was linear, given that the linear model fit best, is undefined because the simpler linear model never fits best. By the same token, the probability that the source model was the threshold process, given that the threshold model fit best, is simply 0.5, or whatever other value is assumed by the unconditional probability of the threshold source. These are Bayesian restatements of the fact that the model with more free parameters will always give a better conventional fit.

In cross-validation, the corresponding probabilities are more informative; Figure 2 shows these probabilities as a function of noise level. Given either source model, the chances of a correct diagnostic decision decline with noise level. The linear source model is identified well at the

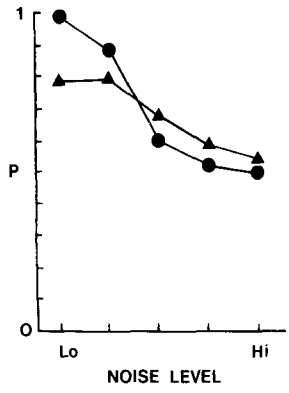

Figure 2. Probability of correctly identifying the source model by choosing the best-fitting model, as a function of source model and noise level. Circles represent the linear source model; triangles represent the threshold source model.

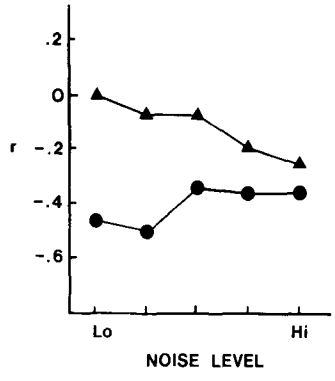

Figure 3. Correlation between the improvement in fit of the threshold model over that of the linear model in conventional analysis of the estimator data set, and the corresponding improvement in cross-validation analysis of the criterion data. A negative correlation means that greater improvements in conventional fit are associated with greater decrements in the quality of cross-validation predictions. Circles represent the linear source model; triangles represent the threshold source model.

lowest noise levels; however, the risk of an incorrect model identification under other conditions is relatively large.

The difference in goodness-of-fit of the two fitted models in the conventional analysis of the estimator data was related to the difference between the two models in their ability to predict the criterion data. Figure 3 shows this relation, expressed as a Pearson *r*, as a function of source model and noise level. The predominantly negative values of the correlation coefficient mean that, in general, a greater margin of conventional goodness-of-fit by the threshold model over the linear model was associated with relatively poorer cross-validation performance by the threshold model. The interaction shown in Figure 3 between source model and noise level may be interpreted as follows. At very low noise levels, the fixed parameters of the threshold source model are very reliably and validly measured by the fitted threshold model. Thus, goodness-of-fit by the threshold model is restricted in range, resulting in a near zero correlation between the threshold model's margins of superiority in the conventional and in the cross-validation analyses.

As the noise level increases, the correlation moves from zero to negative because capitalization on chance tends to maintain the superiority of the threshold fits in conventional model-fitting, at the expense of poorer prediction in cross-validation. Turning to the linear source model, one sees marked negative correlations at all noise levels. These correlations reflect the weakness of a model with many free parameters when it is used to fit data generated by a simpler process. In conventional analysis a weak model will conform to the data closely, but the more closely it fits, the more off-base its predictions will be in cross-validation.

Figure 4 summarizes the main goodness-of-fit patterns observed in conventional model-fitting and in cross-validation. The cross-validation results are from the present study; the conventional results were described earlier in more detail (Collyer, 1985) and confirmed in the present analyses of the present estimator data. In Figure 4, S stands for simple model, represented in this work by the linear model, and C stands for complex model, represented by the threshold model. The figure crosses simple and complex source models, simple and complex fitted models, noise level, and model-fitting procedure, showing how the degree of fit (variance accounted for) behaves under the different conditions.

The traditional ideal and rule of parsimony is illustrated in the upper left panel of Figure 4: In conventional model-fitting, if the degree of fit of two fitted models is very similar, selecting the simpler one will correctly identify the source model; the more complex model deserves to be chosen only when its fit is significantly better. The upper right panel shows that this rule breaks down when

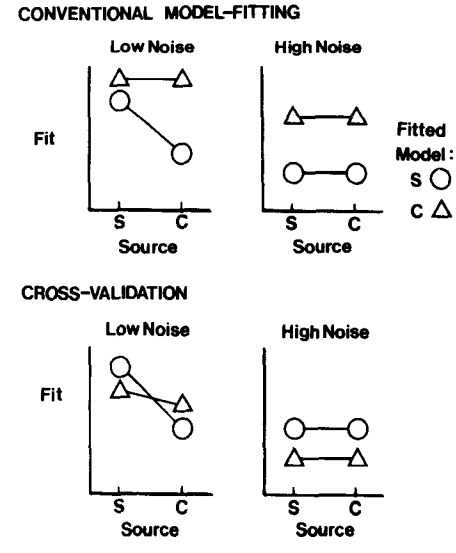

Figure 4. Summary of goodness-of-fit relationships. Circles represent the linear fitted model; triangles represent the threshold fitted model. S = simple model, represented by the linear model. C = complex model, represented by the threshold model.

the data is noisy. When applied to high-noise data, a complex model may capitalize on chance to such an extent that it fits significantly better even when, as a source model, it is false.

The lower left panel illustrates "easy street" for the interpreter of model performance: the best fitting model identifies the source model. It is probable, however, that scientists are not on easy street as often as they would like. It requires very clean data (low noise) and predictions addressed to new data (cross-validation). Just as conventional model-fitting gives complex models an "unfair advantage," cross-validation under high noise has a bias favoring simple models, as shown in the lower right panel of Figure 4. As in the conventional case, sufficiently high noise renders fitted models insensitive to the difference between alternative source models.

CONCLUSION

Cross-validation provides a way of comparing model predictions using *fresh* data, that is, data that has not been used to estimate model parameters. This feature of the procedure may be desirable when the models to be compared have different numbers of parameters, and so capitalize on chance to differing degrees in a conventional model-fitting procedure.

For investigators hoping to find positive evidence for simpler models, it is encouraging that the probability of correct model identification is high at low noise levels. In the data presented here, however, a low noise level is absolutely necessary for a strong case to be made. If possible, it should be shown that the noise level in the data (defined in relation to any models competing for serious consideration) is sufficiently low that the probability of a correct model identification is well above chance.

The models used here to illustrate conventional and cross-validation procedures were developed originally as alternative hypotheses about mental rotation; however, they are not so closely tied to mental rotation as to be irrelevant to other problems. They belong to a class of curve-fitting operations with wide applicability. The question of a threshold arises repeatedly in psychophysics and in other areas of behavioral science, and the special-case relationship between the linear and threshold models reflects a rule rather than an exception in model-based hypothesis testing. For these reasons, a study of these models provides a general methodological perspective on the interpretation of fitted models and the conditions and limitations of source model identification.

Such a study would not be possible without computer simulation and analysis. It did not take a very sophisticated computer to carry out the present study, and, indeed, the models under consideration are not very sophisticated in comparison with other models in science; however, the sheer number of computations necessary to simulate, fit, and correlate hundreds of data sets would have been prohibitive without a computer. More important, the study of model identification requires the ability to control and manipulate known source models. This is easily done through computer programming, whereas in most scientific work, source model identification is the central mystery and the goal of research.

REFERENCES

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal variates. *Annals of Mathematical Statistics*, **29**, 610-611.

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics*, **38**, 476-481.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational & Psychological Measurement*, **11**, 5-11.

Rossi, J. S., & Collyer, C. E. (1986). Is there a threshold for mental rotation? *Bulletin of the Psychonomic Society*, **24**, 1-3.