Biom. J. 32 (1990) 8, 977–983 Akademie-Verlag Berlin

P-Values as Measures of Predictive Validity

G. A. Whitmore

McGill University, Montreal, Canada

E. Xekalaki

The Athens School of Economics and Business Science, Athens, Greece

Abstract

A *predictive P-value* is proposed to measure the difference between an actual and predicted outcome in assessing the validity of an hypothesized prediction model. The concept is illustrated by applications to multiple regression prediction and to the validation of forecast models.

Key words: Forecast; Model; Prediction; Regression; Validation.

1. Introduction

In spite of the concerns of some statisticians about their merit, P-values play a prominent role in statistical analysis and reporting, measuring as they do the discrepancy between an hypothesized model and the statistical evidence bearing on the model's validity. It is proposed here that the role of P-values be expanded to include a *predictive P-value* as a measure of the agreement between a prediction made by a model and the actual outcome. Roughly stated, a predictive P-value is the probability, under the hypothesized prediction model, that the prediction error might have been larger than was actually observed. A small predictive P-value suggests therefore that the prediction model might not be appropriate. A predictive P-value is interpreted in the same manner as P-values in standard hypothesis testing and is based on a similar conceptual framework. It therefore should be of value in statistical education and readily accepted in statistical practice.

2. Theory

For expository convenience, we limit our presentation of the basic theory to univariate predictions. We denote the quantity to be predicted by Y and assume that Y is a continuous random variable that is generated by a true but unknown statistical model M_T. We let $\mathcal{K}$ denote a set of prediction intervals for Y derived from an hypothesized prediction model M_0. The intervals in $\mathcal{K}$ may be random

or fixed, depending on the application. For generality we shall treat them as random. We require $\mathcal{K}$ to have two properties. (1) For each number $c \in (0,1)$ there is an interval $I_c \in \mathcal{K}$. (2) For each $I_c \in \mathcal{K}$, $\Pr\{Y \in I_c\} = c$ if $M_0 \equiv M_T$. The index c denotes the confidence coefficient of interval I_c. The first property requires $\mathcal{K}$ to be a complete set of intervals in the sense of containing an interval for every confidence level. The second property requires that prediction interval I_c have coverage probability c if the hypothesized model M_0 is in fact a true model. The probability statement $\Pr\{Y \in I_c\} = c$ is understood to apply prior to the realization of the interval I_c (if it is random) and the outcome Y. Thus, the coverage probability takes account of the randomness of both Y and I_c.

We define a predictive P-value, denoted here by P, as follows.

$$1 - P = \inf_{\mathcal{K}} \{c \mid Y \in I_c\} . \tag{2.1}$$

Thus, the complement of a predictive P-value is the smallest confidence coefficient among prediction intervals in $\mathcal{K}$ that cover Y.

Prior to realizing the actual outcomes of Y and I_c, the predictive P-value is a random variable that is uniformly distributed on the unit interval if the hypothesized prediction model is a true model, i.e.,

$$P \text{ is distributed as } U(0,1) \text{ if } M_0 \equiv M_T . \tag{2.2}$$

The result holds because, for each $p \in (0,1)$, we have

$$\Pr\{P \leq p\} = \Pr\{\inf_{\mathcal{K}} [c \mid Y \in I_c] \geqq 1 - p\} = p . \tag{2.3}$$

The first equality follows from definition (2.1). The second equality follows from the fact that $\mathcal{K}$ is complete and $\Pr\{Y \in I_c\} = c$ when the hypothesized prediction model is a true model.

The distribution of P when the hypothesized model is not a valid model, i.e., $M_0 \not\equiv M_T$, depends on the nature of the family of prediction intervals $\mathcal{K}$ and on the precise way in which M_T departs from M_0. In most applications, invalid models will tend to give predictive P-values close to 0 or to 1. For example, if M_0 has much greater dispersion than M_T, leading to excessively wide intervals I_c, then P is likely to be large. If M_0 has much less dispersion than M_T or has substantial bias in location, leading to excessively narrow or dislocated intervals, then P is likely to be small. The latter tends to occur frequently in practice. The consistency of predictive P-values against alternative models and their power to detect alternatives is taken up again in the discussion of Section 4.

Prediction results for a model can be aggregated using predictive P-values. For example, if a sequence of predictions by a model A yields the k predictive P-values, $P_1, \ldots, P_k$, then the following geometric mean is a summary descriptive measure of the validity of model A.

$$P_A = \left[\prod_{i=i}^{k} P_i \right]^{1/k} . \tag{2.4}$$

An extreme value of P_A suggests that one or more of the k predictions are inappropriate. For example, if an application yields the five predictive P-values, 0.053, 0.236, 0.146, 0.101, and 0.074, then formula (2.4) gives $P_A = 0.1064$ which is a low P-value.

Predictive P-values derived from the same sample data will have a degree of statistical dependence. Even though the Y outcomes to be predicted in several applications may be independent, if the prediction intervals I_c are calculated from common data then some dependence of the predictive P-values will result. If, however, the P_i are independent then from (2.2) it follows that the P_i constitute a random sample from $U(0,1)$ if M_0 is a true model. In this case, Fisher's method (1932) for aggregating evidence gives the result that

$$(2.5) \qquad -(2k)\ln(P_A) \text{ is distributed as } \chi^2(2k) \text{ if } M_0 \equiv M_T .$$

Thus, for instance, if the five P-values in the preceding numerical example are independent then $k=5$, $P_A = 0.1064$ and, hence, $-(2k)\ln(P_A) = 22.40$. Comparing the latter value with the fractile $\chi^2(0.95; 10) = 18.31$ suggests that one or more of the five predictions of model A are inappropriate.

Again, if $P_1, \ldots, P_k$ are k independent predictive P-values produced in a sequence of predictions by a model, then from (2.2) it follows that a P–P plot or a test of fit to $U(0,1)$ provides a check on the validity of the model. Even if the P-values have some degree of dependence the $P-P$ plot provides a useful approximate check on model validity.

3. Applications

Case 1. Consider the usual multiple regression model in which response variable Y is distributed as $N(X\beta, \sigma^2)$ where X is a vector of k explanatory variables including a constant term and β is the corresponding vector of regression coefficients. Suppose that n independent realizations of the model, (X_i, Y_i), $i=1, \ldots, n$, are available and that it is desired to predict an independent $(n+1)$th realization $Y = Y_{n+1}$ with $X = X_{n+1}$. Following standard procedures [see, for example, NETER, WASSERMAN and KUTNER (1985: 246–247)], a prediction interval for Y having confidence coefficient c takes the form

$$(3.1) \qquad \hat{Y} \pm ts\{Y-\hat{Y}\} ,$$

where $t = t[(1+c)/2; n-k]$ is the $(1+c)/2$ fractile of a $t(n-k)$ distribution, $\hat{Y}$ is the point predictor of Y derived from the fitted model using $X = X_{n+1}$, and $s\{Y-\hat{Y}\}$ is the estimated standard deviation of the prediction error $Y-\hat{Y}$. The predictive P-value associated with the actual outcome Y is the probability number P which solves

$$(3.2) \qquad t[1-(P/2); n-k] = |Y-\hat{Y}|/s\{Y-\hat{Y}\} .$$

As a numerical example, suppose that a multiple regression model is used to

predict crop yield based on weather and other explanatory variables. For a particular application of the model, $n=15$, $k=3$, $\hat{Y}=59.30$ and $s\{Y-\hat{Y}\}=2.28171$. If $Y=65.57$ is the actual crop yield then (3.2) gives

$$t[1-(P/2);\ 12]=(65.57-59.30)/2.28171=2.748$$

and the associated predictive P-value is $P=0.0177$. This P-value is small enough to suggest that the actual outcome might not be consistent with the fitted multiple regression model.

Case 2. The predictive P-value represents a natural performance score for validating forecasting models. It is closely related to scoring rule (3.4) proposed by XEKALAKI and KATTI (1984: 178). These authors give results of two forecast models for corn yields in the years 1963–69 and 1971–80 in two crop reporting districts (CRDs) of the state of Indiana, U.S.A.

Through the courtesy of the authors, the data have been made available to calculate the predictive P-values for these models. The predictive P-values are given in Table 1 and are displayed in overlaid $P-P$ plots in Figure 1. The plotting symbols for the four cases (1, 2, 3, 4) correspond to model A for CRD 20 and CRD 30 and model B for CRD 20 and CRD 30, respectively.

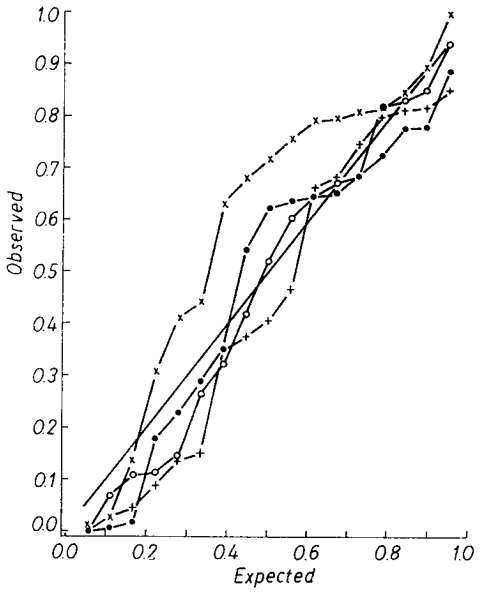
Table 1 shows that the P-values are positively associated across the four cases. The positive association is expected because the two models are predicting the same actual yield in each district and the two districts are in the same state and thus have similar yield patterns over the years. Observe that the actual yields in 1974 (which were very low in both districts) were not predicted well

Table 1

Predictive P-values of crop-yield forecast models A and B for 1963–69 and 1971–80 in two crop reporting districts (CRD 20 and CRD 30) cf the state of Indiana, U.S.A..

	Model A		Model B	
Year	CRD 20	CRD 30	CRD 20	CRD 30
1963	0.747	0.631	0.637	0.520
1964	0.135	0.136	0.229	0.112
1965	0.045	0.026	0.017	0.069
1966	0.683	0.757	0.652	0.939
1967	0.149	0.815	0.885	0.685
1968	0.375	0.308	0.623	0.108
1969	0.813	0.846	0.288	0.603
1971	0.348	0.795	0.355	0.853
1972	0.662	0.679	0.541	0.323
1973	0.406	0.716	0.778	0.831
1974	0.000	0.012	0.000	0.000
1975	0.852	0.808	0.646	0.419
1976	0.087	0.442	0.177	0.645
1977	0.819	0.413	0.780	0.145
1978	0.468	0.895	0.687	0.672
1979	0.800	0.791	0.725	0.266
1980	0.030	1.000	0.006	0.820

by either model. The $P-P$ plots show that the hypothesized predictive distributions tend to have (1) thin tails in all four cases, resulting in too many small P-values, and (2) low kurtosis in cases 2 and 3 (model A in CRD 30 and model B in CRD 20), resulting in too many P-values in the upper middle range of the (0,1) interval.



Plotting Symbols

1. +—+—+
2. ×—×—×
3. ·—·—·
4. ○—○—○

Fig. 1. Overlaid $P-P$ plots of the predictive P-values in Table 1. The plotting symbols (1, 2, 3, 4 correspond to model A for CRD 20 and CRD 30 and model B for CRD 20 and CRD 30, respectively.

Case 3. Although the regression prediction setting is an important one for predictive P-values, the concept can be more widely applied. Suppose, for example, that M_0 specifies that Y has the probability density function $f_0(Y)$. Let the prediction intervals I_c in $\mathfrak{X}$ be defined by

$$I_c = \{y \mid f_0(y) \geqq r\} \tag{3.3}$$

where r is chosen so

$$\int_{I_c} f_0(y)\, dy = c\ .$$

Since $f_0(y)$ is given, the intervals in this case are not random. A predictive P-value can be calculated using (2.1) to check the agreement of the outcome of Y with the hypothesized model $f_0(y)$. For example, if $f_0(y)$ is the normal density function corresponding to $N(80, 100)$ then the prediction intervals defined by (3.3) have the form $80 \pm 10z$ where $c = \Pr\{|Z| \leqq |z|\}$ and Z is the standard normal deviate. If the actual outcome is, say, $Y = 93$, then $z = (93-80)/10 = 1.30$ and the associated predictive P-value is $P = 1 - \Pr\{|Z| \leqq 1.30\} = 0.1936$. The moderately large value of P here suggests that the actual outcome is in fair agreement with the hypothesized model.

4. Discussion

The concept of a predictive P-value may be extended in a number of ways. For example, the use of a predictive P-value is not restricted to a univariate outcome. It can be computed from definition (2.1) for cases where the predicted outcome Y is multivariate and I_c is a multidimensional prediction region. The concept also extends readily to situations where a set of simultaneous predictions of Y are to be made, as for example, with simultaneous predictions of response variable Y at different levels of the explanatory variables X in a multiple regression application.

The predictive P-values calculated from an hypothesized model reflect the *validity* of the model but not its predictive *accuracy*. For example, the normal regression model described in Case 1 above will yield predictive P-values that are uniformly distributed on (0,1) if it is a true model for the particular application of interest. Likewise, however, a predictive model that is based simply on the corresponding marginal distribution of response variable Y will also yield uniformly distributed predictive P-values in the same application. Both models are valid but the first will be more accurate if the explanatory variables X have any degree of predictive power.

As noted earlier, the consistency and power of predictive P-values in detecting alternatives to the null model M_0 is a practical concern. We now discuss this issue briefly.

We have not insisted in our theoretical development that the set of prediction intervals $\mathfrak{I}$ be constructed in any way that would optimize their power to reject an invalid model. Predictive P-values can be calculated whether $\mathfrak{I}$ has optimal properties or not. We note, however, that the set of prediction intervals can be constructed so that they are consistent against all alternative models. To demonstrate this claim, suppose that Y has support on the whole real line and the intervals in $\mathfrak{I}$ are the set of half-lines $(-\infty, u)$ where u ranges over all real numbers. In this case, the predictive P-value has a uniform distribution if and only if M_0 is a true model. This conclusion follows from the result in (4.1) below, in which F_0 and F_1 denote the cumulative distribution functions of Y under the null and alternative hypotheses, respectively.

$$\begin{aligned}(4.1)\qquad \Pr\{P \leqq p\} &= \Pr\{\inf_{\mathfrak{I}} [c \mid Y \in I_c] \geqq 1-p\} \\ &= \Pr\{\inf_{u \in R} [F_0(u) \mid Y \in (-\infty, u)] \geqq 1-p\} = F_1[F_0^{-1}(p)]\,.\end{aligned}$$

Since F_1 and F_0 are continuous cumulative distribution functions, it follows that the right-most expression in (4.1) can be identical to p for all $p \in (0,1)$ if and only if $F_1 \equiv F_0$.

On the other hand, there are sets of prediction intervals $\mathfrak{I}$ that are optimal in several respects and yet are not consistent against all alternative models. In (3.3), for example, suppose that $f_0(y)$ denotes the standard normal density

function. Then the set of intervals $\mathfrak{K}$ defined by (3.3) in this case are of the form $(-z, z)$ for all $z \in (0, \infty)$. Let $h(y)$ be any odd function of y having the property that

$$f_1(y) = f_0(y) + h(y) \geqq 0 \quad \text{for all } y \,. \tag{4.2}$$

For instance, $h(y) = f_0(y) \sin(y)$ is such a function. It is readily shown that $f_1(y)$, defined in this manner, represents a proper probability density function. Moreover, predictive P-values calculated from the intervals in $\mathfrak{K}$ in this case will be uniformly distributed whether Y is distributed according to $f_0(y)$ or $f_1(y)$. Although the intervals in $\mathfrak{K}$ here are not consistent against this peculiar type of alternative model, set $\mathfrak{K}$ is consistent and powerful against a whole range of plausible alternatives. For example, if the plausible alternatives consist only of symmetrically distributed random variables, set $\mathfrak{K}$ will perform very well.

Thus, to achieve consistency against plausible alternative models and, moreover, to achieve reasonable power, the type of prediction interval employed must be chosen with some care. Thereafter, the calculation of predictive P-values can prove effective in assessing the validity of prediction models in comparison with plausible alternative models.

Acknowledgement

This research was partially funded by a grant (to Whitmore) from the Natural Sciences and Engineering Research Council of Canada.

References

FISHER, R. A., 1932: *Statistical Models for Research Workers,* 4th edition, London: Oliver and Boyd.

NETER, J., W. WASSERMAN and M. H. KUTNER, 1985: *Applied Linear Statistical Models,* 2nd edition, Homewood, Illinois: Irwin.

XEKALAKI, E. and S. K. KATTI, 1984: A technique for evaluating forecasting models. Biom. J. **26**, 173–184.

Received July 1989

Prof. G. A. WHITMORE
Management Science
Faculty of Management
McGill University
1001 Sherbrooke Street West
Montreal, Quebec H3A 1G5
Canada

Prof. E. XEKALAKI
The Athens School of
Economics and Business Science
Athens, Greece