A MEASURE OF PREDICTIVE PERFORMANCE FOR ESTIMATION
WHEN THE PREDICTED VALUE IS A PROBABILITY

Josefina A. Lago, Westat, Inc.*

ABSTRACT

Traditionally, the R² statistic is used as the measure of predictive performance. However, when the estimation method is non-linear, the R² loses some of its desirable properties.

This paper presents an alternative measure of predictive performance, Relative Information (IR), developed by Betancourt and Clague and modified by Lago.

The IR measure is suitable for non-linear models where the dependent variable is categorical and the predicted values can be interpreted as probabilities. MACRO RELINFO, developed to compute the IR statistic, requires as input for each observation:

(a) the observed response,
(b) the estimated probabilities associated with each value of the response variable.

The output from the macro is the value of the IR statistic and selected intermediate steps.

INTRODUCTION

In recent years, there has been an increased interest in models suitable for categorical dependent variables. Non-linear estimation methods that map a vector of explanatory variables X into the zero-one interval have been used most successfully to relate a set of explanatory variables to a qualitative dependent variable. However, the traditional measure of predictive performance, R², loses some of its desirable properties when the estimation method is non-linear; for instance, the usual decomposition of total variation is not applicable.

This paper discusses an alternative measure of predictive performance, Relative Information, (IR), suitable in non-linear estimation situations where the estimated values can be interpreted as probabilities. A description of the computation of IR is presented and an example illustrates the use of MACRO RELINFO, developed to compute the IR statistic.

THE RELATIVE INFORMATION-IR-STATISTIC

Let the categorical dependent variable Y represent K possible responses with probabilities of occurrence P₁, P₂,...,Pₖ. In general, the models to which IR is applicable postulate that the probability associated with the ith response, Pᵢ, is a non-linear function of some vector X of explanatory variables, \( f(X'B) \), where \( B \) is a vector of unknown coefficients.

The measure of predictive performance IR is based on the concept of entropy, which may be interpreted as a measure of the amount of uncertainty associated with the distribution of \( \{P₁, P₂,...,Pₖ\} \). Entropy, \( Eᵢ \), is defined for the ith observation by:

\[
Eᵢ = - \sum_{k=1}^{K} P_k \log P_k \quad (1)
\]

For a dichotomous response variable, such as whether or not a household invests in storm windows or doors,

\[
P_{i1} = P(\text{household } i \text{ invests in storm windows or doors})
\]

\[
P_{i2} = P(\text{household } i \text{ does not invest})
\]

Thus, the value of entropy for the ith observation is given by:

\[
Eᵢ = - \left[ \hat{P}_{i1} \log \hat{P}_{i1} + (1-\hat{P}_{i1}) \log (1-\hat{P}_{i1}) \right] \quad (2)
\]

where \( \hat{P}_{i1} \) and \( \hat{P}_{i2} \) are the model estimates of \( P_{i1} \) and \( P_{i2} \) defined above.

Clearly, the greatest amount of uncertainty will be associated with \( P_k = 1/K \) for all \( k \), and the minimum at \( P_k = 1 \) for any \( k \) (\( Eᵢ = 0 \)). Letting \( E_{\text{max}} \) be the maximum amount of entropy associated with the distribution, i.e.,

\[
E_{\text{max}} = - \sum_{k=1}^{K} \frac{1}{K} \log (1/K) = - \log (1/K) \quad (3)
\]

Then, a measure of the amount of information contained in the predicted probabilities \( P_k \) (\( k=1,...,K \)) can be obtained as a function of the entropy \( Eᵢ \) by

\[
Iᵢ = 1 - \left( \frac{Eᵢ}{E_{\text{max}}} \right) \quad (4)
\]

Betancourt and Clague define a "correct" prediction as one for which \( \hat{P}_k \geq 1/K \), when the observed response is \( k \), and \( \hat{P}_k < 1/K \) for other values of the response variable. However, when the response variable has more than three values, it is difficult to obtain a "correct" prediction under this definition.
Alternatively, a less demanding definition would consider a predicted probability "correct" if \( P_k > P_h \) \( h \neq k \) when \( k \) is the observed response, and "incorrect" otherwise. This definition of "correct" is the one employed in MACRO RELINFO.

The amount of information contained in a set of \( N \) predictions may be defined as:

\[
\bar{I} = (I_C - I_I)/N \tag{5}
\]

where \( I_C \) is the sum of information \( I_i \) for all the "correct" predictions, and \( I_I \) is the sum of all "incorrect" predictions. \( I \) ranges from -1 (all probabilities incorrectly predicted) to +1 (all probabilities correctly predicted). It should be noted that more credit (discredit) is given to a correct (incorrect) prediction if a high probability underlies the prediction, i.e., close to 1 or 0, than a low one, i.e., close to 1/K.

When the distribution of the observations over the possible response values is very uneven, there is a significant amount of information contained in the "naive" predictions given by the sample proportions associated with each response. On the other hand, if the proportion of observations in the sample at each response value are evenly divided, the amount of information from the "naive" predictions will be relatively low. To take into account these distinct situations, Betancourt and Claque suggest using an adjustment to \( \bar{I} \) which captures the absolute amount of additional information provided by the introduction of the theory over the information already contained in the sample proportions. Thus,

\[
I_A = \bar{I} - \bar{I}_M \tag{6}
\]

where \( \bar{I} \) is defined in (5) and \( \bar{I}_M \) is obtained by substituting the sample proportions associated with each response in (1), instead of \( P_k \), and substituting \( \bar{P}_M \) in (4) to obtain \( \bar{I}_M \). Then a measure of the amount of information provided by the introduction of the theory relative to the maximum amount of information which the theory can capture in a given sample, is given by

\[
I_R = I_R/(1 - \bar{I}_M) \tag{7}
\]

In practice, the value of \( I_R \) is rarely greater than .3. As long as \( I_R \) is positive, this indicates that the model is contributing to explaining the variability in the data, after adjusting for incorrect predictions and the information contained in the sample proportions. Furthermore, it should be noted that \( I_R \) does not necessarily increase with the number of explanatory variables included in the model, as is the case with \( R^2 \).

A major advantage of the measure of relative information \( I_R \) is that it is equally applicable to situations where the response variable has more than two values; in contrast, the \( R^2 \) cannot be defined in those cases.

MACRO RELINFO

To illustrate the use of MACRO RELINFO, assume the states have the choice of selecting one of three energy conservation programs to be funded by the Federal Government:

- Industrial Energy Audits;
- Commercial Energy Audits;
- Residential Energy Audits.

A multinomial logit model postulates that the state's choice among the three programs is a function of characteristics such as the state's population, value added of manufacturing, employment in selected services, etc. Table 1 presents the estimated probabilities associated with each possible choice. These probabilities constitute the input data set A to MACRO RELINFO.

The proportion of states that choose each conservation program is obtained by PROC FREQ. The results are outputted to data set B, which is given in Table 2.

Finally, Table 3 presents a sample run of the MACRO and the results obtained when using the data sets A and B described above.

*This research was started while employed at the U.S. Department of Agriculture.

REFERENCES

1. R. Betancourt and C. Claque, Capital Utilization: A Theoretical and Empirical Analyses. (Book manuscript accepted by Cambridge University Press).

Table 1. Estimated Probabilities

LISTING OF DATA SET A: COLUMN 1 IS VALUE OF RESPONSE VARIABLE
COLUMNS 2 TO 4 ARE ESTIMATED PROBABILITIES ASSOCIATED WITH EACH RESPONSE

<table>
<thead>
<tr>
<th>OBS</th>
<th>RESPONSE</th>
<th>INDUSTRI</th>
<th>COMERCI</th>
<th>RESIDENT</th>
<th>RESIDENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.46</td>
<td>0.13</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.39</td>
<td>0.42</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.42</td>
<td>0.20</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.28</td>
<td>0.29</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.68</td>
<td>0.21</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.81</td>
<td>0.06</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.27</td>
<td>0.32</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.52</td>
<td>0.19</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.37</td>
<td>0.28</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.49</td>
<td>0.04</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.63</td>
<td>0.15</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.69</td>
<td>0.26</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.51</td>
<td>0.39</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0.38</td>
<td>0.26</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0.62</td>
<td>0.21</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Data Set B

SAMPLE PROBABILITIES OF EACH RESPONSE
LISTING OF DATA SET SAMPLE

<table>
<thead>
<tr>
<th>OBS</th>
<th>RESPONSE</th>
<th>COMP. PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.3546167</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.2291687</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.468867</td>
</tr>
</tbody>
</table>

425
Table 3. Sample Run

```
OPTIONS VNOTES NODATE NONUMBER!

MACRO HELINFO COMPUTES THE RELATIVE INFORMATION (1X) &;
MEASURE OF PREDICTIVE PERFORMANCE WHICH IS SUITABLE FOR &;
ESTIMATION SITUATIONS WHERE THE PREDICTED VALUES CAN BE &;
INTERPRETED AS PROBABILITIES.

WHEN THE RESPONSE VARIABLE IS MISSING THE WHOLE OBSERVATION &;
SHOULD BE DELETED ALSO WHEN THE ESTIMATED PROBABILITIES &;
ARE ALL ZERO OR ONE OR NONE OF THE ESTIMATED PROBABILITIES &;
ARE MISSING THE WHOLE OBSERVATION SHOULD BE DELETED.
THE FOLLOWING STATEMENTS ARE REQUIRED TO EXECUTE MACRO HELINFO:

WHERE A IS THE NAME OF THE DATA SET CONTAINING &;
THE VALUE OF THE RESPONSE VARIABLE (POSITION 1) &;
AND THE ESTIMATED PROBABILITIES (POSITION 2).&;
D IS THE NAME OF THE DATA SET CONTAINING THE &;
SAMPLE PROPORTIONS (POSITION 2) ASSOCIATED WITH EACH &;
LEVEL OF THE RESPONSE VARIABLE (POSITION 1).&;

NOTE THAT DATA A CAN BE AN OUTPUT DATA SET FROM PHUG FREQ:
PHUG FREQ:
TABLES RESPONSE/OUT=A:
DATA B:
SET B:
SAMP.PROCENT=10 WHERE:
KEEP RESPONSE SAMP.HH:

MACRO USNAME &;
MACRO SAMPNAME &;
MACRO HELINFO:

WHERE A IS THE NAME OF THE DATA SET CONTAINING &;
THE VALUE OF THE RESPONSE VARIABLE (POSITION 1) &;
AND THE ESTIMATED PROBABILITIES (POSITION 2).&;
D IS THE NAME OF THE DATA SET CONTAINING THE &;
SAMPLE PROPORTIONS (POSITION 2) ASSOCIATED WITH EACH &;
LEVEL OF THE RESPONSE VARIABLE (POSITION 1).&;

PHUG FREQ:
TABLES RESPONSE/OUT=A:
DATA B:
SET B:
SAMP.PROCENT=10 WHERE:
KEEP RESPONSE SAMP.HH:

MACKO HELINFO:
MACRO M vents:
MACRO M vents:
MACRO HELINFO:

---DEFINE MACTRICES AND DIMENSIONS---
N=INROW(S):
SAMP=INROW(FROMS):
N=INROW(FROMS):
SAMP=INROW(FROMS):
N=INROW(FROMS):
SAMP=INROW(FROMS):

---RECODE 0-1 DEPENDENT VARIABLES---
If MAX(NRESP)=1 AND MIN(NRESP)=0 THEN DO:
INO=1 TO NUMS:
If NRESP(i)=1 THEN NRESP(i)=2:
If NRESP(i)=2 THEN NRESP(i)=1:
END:

---PRINT ERROR MESSAGE WHEN DEPENDENT VARIABLE OR ESTIMATED &;
PROBABILITIES ARE MISSING ---

NO=1 TO NUMS:
If NRESP(i)=0 THEN PRINT:
Note missing dependent variable for following obs:
PRINT:
STOP:
STOP:
END:
```

Table 3. (continued)

```plaintext
IF ESTPHS(I,*)=ZEMU(*) THEN DO;
  PRINT I;
  STOP;
END;

IF ANY(ESTPHS(I,*)#MISS(*)) THEN DO;
  NOTE ESTIMATED PROBABILITIES MISSING FOR FOLLOWING OBS;
  PRINT I;
  STOP I;
END;

END;

/*--COMPUTE MAXIMUM ENTROPY--*/
IF MAX(ESTPHS(I,K))=ESTPHS(I,*) THEN 
  CORRECT=CORRECT+INF0;
ELSE 
  INCOR=INCOR+INF0;
END;

/*--COMPUTE ENTROPY--*/
ENT0U0=
DO I=1 TO N081
  ENT=ENT+ESTPHS(I,K)*LOG(ESTPHS(I,K))
END;

/*--COMPUTE INFORMATION MEASURE--*/
INF0U0=ENT/#EMAX
IF ESTPHS(I,RESPI(I))=MAX(ESTPHS(I,*)) THEN 
  CORRECT=CORRECT+INF0;
ELSE 
  INCOR=INCOR+INF0;
END;

/*--COMPUTE ENTROPY AND INF0 USING PROBABILITIES BASED ON SAMPLE PROPORTIONS--*/
ENT0U0=
DO I=1 TO N081
  ENT=ENT+SAMPRI(*L)*LOG(SAMPRSl(*L))
END;

/*--COMPUTE MEAN IM--*/
IMBAR=(MCUR+MINC)/#N081
IREL=(IBAR-IMBAR)/#N081
*/

/*--OUTPUT RESULTS AND PRINT STATEMENT--*/
OUTPUT [BAR OUT=RESULTI(H8NAME=(COL1=IBAR))]
OUTPUT IMBAR OUT=RESULT2(R8NAME=(COL1=IMBAR))
OUTPUT IREL OUT=RESULT3(R8NAME=(COL1=IREL))
DATA RESULTI : MERGE RESULTI RESULT2 RESULT3
FILE PRINT NOTITLE:
  PUT W25 : THE MEASURE OF IMBAR = I IMBAR 10.6 ///
  PUT W25 : THE VALUE OF IMBAR = I IMBAR 10.6 ///
  PUT W25 : FINAL RESULT ///
  /*--DELETE THE FILE PRINT STATEMENT FOR INTERACTIVE USE--*/
  &
```

427
Table 3. (continued)

DATA STATES: INPUT RESPONSE INSTRNL COMERCL RESIONL

CARUS;

PROC PRINT;
TITLE CONSERVATION PROGRAM SELECTED BY STATE AND PREDICTED PROBABILITIES;
TITLE1 LISTING OF DATA SET A: COLUMN 1 IS VALUE OF RESPONSE VARIABLE;
TITLE2 COLUMNS 2 TO 4 ARE ESTIMATED PROBABILITIES ASSOCIATED WITH EACH RESPONSE;
TITLE3;

PROC FREQ; TABLES RESPONSE/ OUT=SAMPLE1;

DATA SAMPLE1; SET SAMPLE1;
SAMP_PH=.01*PERCENT;
KEEP RESPONSE SAMP_PH;

PROC PRINT;
TITLE;
TITLE2 SAMPLE PROBABILITIES OF EACH RESPONSE;
TITLE3 LISTING OF DATA SET SAMPLE1;
TITLE4;
*--INVOKE MACRO RELINFO--;
MACRO PHSOIS STATES *
MACRO SAMPSIN SAMPLE *

RELINFO

INTERMEDIATE RESULTS

THE VALUE OF IBAR = 0.175347

THE VALUE OF IMBAR = -0.004336

FINAL RESULT

THE MEASURE OF RELATIVE INFORMATION (IN) EQUALS 0.178907

428