THE COEFFICIENT OF PREDICTION FOR MODEL SPECIFICATION

Frank G. Landram, West Texas A & M University
Amjad Abdullat, West Texas A & M University
Vivek Shah, Southwest Texas State University

ABSTRACT

The coefficient of prediction P_j^2 is derived from the PRESS (prediction sum of squares) statistic just as R_j^2 is derived from SSE, the error sum of squares. While R_j^2 measures quality of fit, P_j^2 measures quality of point predictions. Unlike SSE and PRESS, R_j^2 and P_j^2 are bounded, relative measures ideally suited for statistical modeling. This paper describes the limits, properties, how P_j^2 differs from other criteria, and the rationale for its importance. This knowledge enhances one's understanding of what constitutes properly specified statistical models. An example illustrates the behavior and practical applications of P_j^2 in model specification analysis.

INTRODUCTION

Consider the sample least squares equation,

$$\hat{\mathbf{Y}} = \mathbf{Xb}, \qquad (1)$$

where **X** is a data matrix of full rank, **b** is a vector of regression coefficients, and $\hat{\mathbf{Y}}$ is a vector of fitted values. An evaluation of (1) is obtained from the following statistical modeling criteria:

(a) Goodness of fit criteria include measures such as R_j^2 and adjusted $\overline{R}_j^2$ (Haitovsky, 1969).

(b) Smallness of prediction variances criteria (the mean square errors of prediction) include measures such as Mallows C_p (1973), Akaika AIC (1969), and Amemiya PC_j (1980).

c) Goodness of prediction criteria include measures such as "split data analysis," PRESS (Hocking, 2003), and P_j^2.

The subscript j notation denotes the number of X-variables in the model.

Observe that quality of predictions is measured by two criteria. The statistics

in (b) are primary concerned with the inflation of prediction variances (Maddala, 2001). The statistics in (c) are confined to measuring the accuracy of out-of-sample point predictions. Although related, these statistics sometime behave differently than the statistics in (b). An example illustrates the need for both criteria. It illustrates that the inclusion of a multicollinear variable may diminish the accuracy of point predictions while improving interval predictions by reducing the size of their prediction variances. Hence, both criteria are needed in the proper specification of a statistical model. This example also proves the contrary to statements found in textbooks that multicollinearity is not harmful to point predictions. Several authors realize the importance of P_j^2 and are currently using the measure in validation and statistical modeling. (Myers, 1990, also Montgomery, Peck and Vining, 2001). Regrettably, the literature is void of the statistical properties and limits of P_j^2. Therefore, it seems noteworthy to describe these properties, how P_j^2 differs from other criteria, and the rationale for its importance in statistical modeling.

THE P^2 STATISTIC

Let an out-of-sample prediction $\hat{Y}_{(i)}$ be computed by using a "new" observation in (1). Since Y_i of the holdout observation is not used in fitting the regression model, the out-of-sample predicted value $\hat{Y}_{(i)}$ is independent of Y_i in calculating the PRESS residual

$$e_{(i)} = Y_i - \hat{Y}_{(i)} \qquad (2)$$

This "leave one out" process is repeated n times. Computationally, PRESS residuals in (2) are obtained from

$$e_{(i)} = e_i /(1-h_{ii}), \qquad (3)$$

where e_i are ordinary least squares residuals from (1) and h_{ii} are the diagonal elements of the hat matrix. (Hoaglin and Welsch, 1978). Observe, PRESS residuals are weighted least squares residuals with $1/(1-h_{ii})$ being the weights.

Properties of h_{ii}. The diagonal elements h_{ii} in (3) possess the following properties:

$$1/n \leq h_{ii} \leq 1.0 \qquad \text{and} \qquad \sum_{i=1}^{n} h_{ii} = p$$

(Belsley, Kuh, and Welsch, 1980). All diagonal h_{ii} values are between 1/n and one, given that (1) has an intercept. Also, the sum of the diagonal h_{ii} values equals p, the number of regression coefficients in (1).

Hat Matrix. The hat matrix is an extremely efficient n x n projection matrix defined as

$$\mathbf{[h_{ij}] = H = X(X'X)^{-1}X';} \qquad (4)$$

where **X** is an n x *p* data matrix of full rank and $\mathbf{(X'X)^{-1}}$ is the traditional least squares inverse matrix. Knowing that the vector **b** in (1) is

$$\mathbf{b} = \mathbf{(X'X)^{-1}X'Y} \tag{5}$$

then (1) can be written as

$$\hat{\mathbf{Y}} = \mathbf{HY} \tag{6}$$

It is important to know that the hat matrix in (4) is a projection matrix. Therefore, its diagonal elements h_{ii} usually increase (can never decrease) when an additional variable enters the model (Hoaglin and Welsch, 1978).

PRESS. The PRESS statistic (Walls and Weeks, 1969) is the sum of the squared PRESS residuals;

$$\text{PRESS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_{(i)})^2 = \sum_{i=1}^{n} e_{(i)}^2 \tag{7}$$

The independence of Y_i and $\hat{Y}_{(i)}$ in (7) enables the PRESS statistic to be a true assessment of the validity or prediction capabilities of the regression model. PRESS statistics are similar to the error sum of squares in regression analysis,

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{8}$$

While SSE uses fitted values $\hat{Y}_i$, PRESS in (7) uses out-of-sample predicted values $\hat{Y}_{(i)}$. Hence, just as SSE is used in calculating the coefficient of determination,

$$R_j^2 = 1 - (\text{SSE} / \text{SST}), \tag{9}$$

PRESS is used in calculating the coefficient of prediction,

$$P_j^2 = 1 - (\text{PRESS} / \text{SST}); \tag{10}$$

where $\text{SST} = \Sigma(Y_i - \overline{Y})^2$. By dividing the PRESS statistic by SST and subtracting the ratio from one, P_j^2 is similar to R_j^2 both are relative measures.

The Behavior Of P_j^2.

The PRESS residuals $e_{(i)}$ in (3) are a function of least squares residuals e_i and diagonal elements h_{ii} of the hat matrix. Since the hat matrix is a projection matrix, its diagonal elements h_{ii} usually increase (never decrease) when additional variables enter the model (Myers, 1990). Let k = p+1, then an incoming variable will cause $\text{PRESS}_k \leq \text{PRESS}_p$ and $P_k^2 \geq P_p^2$ only if the least squares residuals e_i in (3) decrease

proportionally more than the increase in the weights $1/(1-h_{ii})$. Therefore, the P_j^2 criterion incurs penalties for including irrelevant variables and for deleting relevant variables. These penalties are assessed differently than penalties assessed by other model specification criteria.

Irrelevant Variables. A decrease in P_j^2 signals the inclusion of an irrelevant variable and that the model is becoming overspecified. Hence, the inclusion of a multicollinear variable may cause the accuracy of point predictions to diminish. This concept is illustrated in the following example and contradicts statements in many textbooks that multicollinearity is not harmful to point predictions.

Limits of P_j^2. Observe that out-of-sample observations are not used in deriving the statistical model. Therefore, independent, out-of-sample predictions $\hat{Y}_{(i)}$ are not as accurate in predicting Y_i as in-sample fitted values. Hence, P_j^2 can not exceed the expected value of the determination coefficient R_j^2, adj $\overline{R}_j^2$ and Amemiya's PC_j. Also, since $e_{(i)} = e_i/(1-h_{ii})$, PRESS residuals in (3) are weighted least squares residuals causing $e_{(i)} > e_i$. This, in turn, causes $PRESS_j > SSE_j$, and $P_j^2 < R_j^2$. Again, the expected accuracy of out-of-sample predictions measured by P_j^2 cannot exceed the accuracy of in-sample estimates.

HOSPITAL STAFFING EXAMPLE

This example illustrate that P_j^2 may behave differently from other modeling criteria. Monthly labor hours (Y_i) for 17 U.S. Hospitals are analyzed using data from Navy Manpower and Material Center, 1979. X_1 is average daily patient load; X_2 is monthly X-ray exposures; X_3 is monthly occupied bed days; X_4 is eligible population in area X_5 is average number of days a patient stays in the hospital. By employing the all possible regression algorithm, the various subsets of the five variable model are analyzed. Equations possessing the maximum value for the criteria in each subset are recorded in Table 1. In addition to R_j^2 and P_j^2 other criteria frequently used in statistical modeling are the adjusted $\overline{R}_j^2$ Mallows C_p, and Amemiya's PC_j criteria. After disregarding Mallows C_p (it is not in a relative-comparable form), Figure 1 reveals that the above criteria are larger than P_j^2. This is expected since in-sample residuals are smaller than out-of-sample residuals. Notice that within a specific subset, maximum P_j^2 may not belong to the same equation as the other criteria.

Table 1
Statistical Modeling Criteria

Model	R_j^2	$\overline{R}_j^2$	C_p	PC_j	P_j^2	variables
6	0.9908	0.9867	6.000	0.9808	0.9349	X_1 X_2 X_3 X_4 X_5
5	0.9908	0.9877	4.026	0.9832	0.9421	X_2 X_3 X_4 X_5
5	0.9851	0.9801	10.922	0.9726	0.9624	X_1 X_3 X_4 X_5
4	0.9901	0.9878	2.918	0.9840	0.9639	X_2 X_3 X_5
4	0.9850	0.9816	8.968	0.9758	0.9736	X_1 X_3 X_5
3	0.9867	0.9848	4.942	0.9810	0.9639	X_2 X_3
3	0.9848	0.9826	7.294	0.9782	0.9745	X_3 X_5
2	0.9722	0.9703	20.381	0.9648	0.9559	X_2

Except for P_j^2 all of the above criteria are a function of the mean square error, MSE (Maddala, 2001). Therefore, Table 1 reveals that within each subset, the equation possessing the largest R_j^2 also possesses the largest $\overline{R}_j^2$ and PC_j and smallest C_p. This is often the case. However, it is not necessary true for P_j^2. Equation $Y = f(X_2, X_3, X_5)$ possesses the global maximum/ minimum for all criteria except P_j^2 and R_j^2. Since R_j^2 is upward biased (never decreases), its maximum is not considered. Observe, the global maximum for P_j^2 is located in the two variable subset. Also observe that $P_j^2 = 0.9639$ for models $Y = f(X_2, X_3)$ and $Y=f(X_2, X_3, X_5)$. Thus, X_5 adds nothing to the accuracy of point predictions for these models. Figure 1 illustrates that the importance of additional variables diminish as more variables are added to the model. The general configuration of the P_j^2 curve is always cupped downward--given that the full equation is over specified.

DISCUSSION

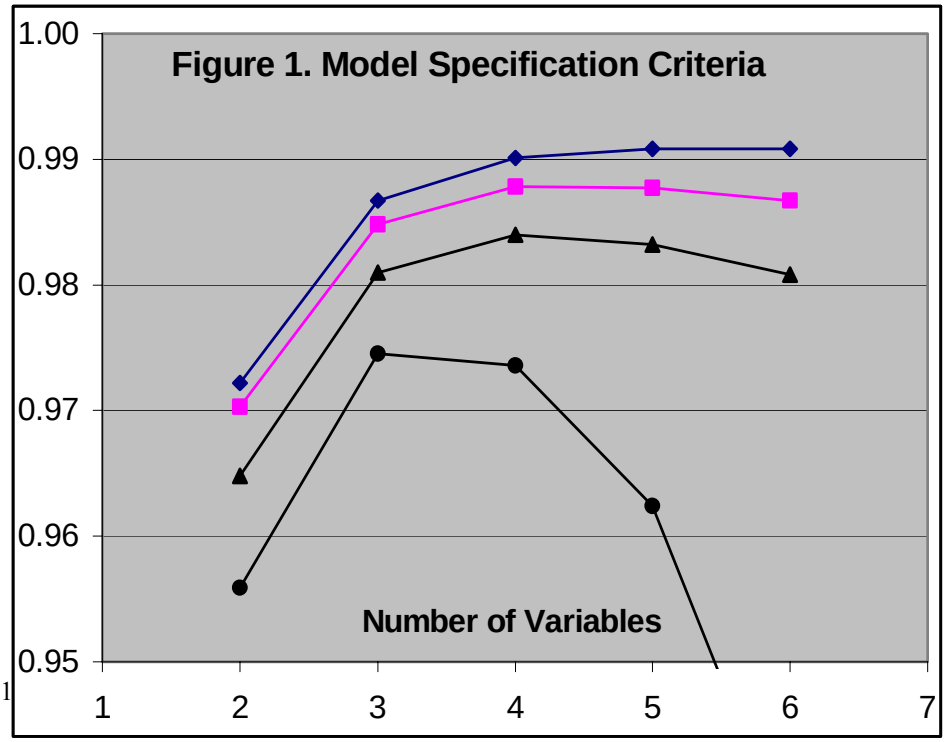
Leverage. As revealed by (3), (7), and (10), P_j^2 is a function of least squares residuals e_i and the diagonals elements h_{ii}. These elements are called leverage values and often cause P_j^2 to behave in a manner different from other criteria. Large leverage h_{ii} values are caused by extreme X-values in the i^{th} observation and indicate potentially strong influence on the statistical model (Maddala, 2001). Since these h_{ii} values are used as

weights in PRESS residuals, observations that have a strong influence on predictions are reflected in P_j^2. When the global maximum P_j^2 coefficient does not agree with the other statistical modeling criteria, one should investigate possible model underspecification. In the above example, the last four observations are from large hospital installations (over 10,000 hours); some reveal unusually large PRESS residuals. By including a dummy variable in the equation, the global maximum / minimum for all criteria (including P_j^2) focuses on two models:

Model	R_j^2	$\overline{R}_j^2$	C_p	PC_j	P_j^2
$Y = f(X_1, X_2, X_5, D)$	0.9968	0.9957	3.602	0.9941	0.9906
$Y = f(X_2, X_3, X_5, D)$	0.9968	0.9957	3.533	0.9941	0.9894

where D = 1 if Y > 10,000, zero otherwise. The significance of the dummy variable separates large hospitals into a separate class for more accurate predictions. Observe, the global maximum for P_j^2 goes from 0.9745 in Table 1 to 0.9906. Interaction variables are insignificant suggesting that large hospitals respond to changes in the X-values at the same rate as small hospitals. Of course, knowledge of the subject matter must always be used in selecting the most appropriate model.

Figure 1
Model Specification Criteria



CONCLUDING REMARKS

Information concerning model specification are abundant in the literature. Additional criteria that are used in selecting possible regression models can be found in Hocking (1976) and Akaike (1969). A minimum mean square error of prediction or a maximum quality of fit does not guarantee maximum accuracy for point predictions. Although related, the accuracy of point predictions and the size of interval predictions do not necessary act in one accord. This is evident from the behavior of P_j^2 in Table 1. An incoming variable may increase the accuracy of point predictions while inflating interval predictions; as confirmed in Table 1, the reverse is also true. Therefore, a statistic confined to measuring the accuracy of point predictions also needs consideration. Hence, this study provides a through understanding of what constitutes properly specified statistical models.

Unlike the PRESS statistics, the P_j^2 statistic is a bounded, relative measure of prediction that can be directly compared with other statistical modeling criteria. This statistic utilizes weighted least squares residuals in minimizing the sum of the squared PRESS residuals. Penalties for including irrelevant variables and for deleting relevant variables are associated with this statistic. Observe that $\overline{R}_j^2$ tends to minimize the mean square error of fitted observations. Ameniya's PC_j and Mallows C_p criteria strive to minimize the mean square error of prediction. The P_j^2 criterion strives to minimize PRESS residuals thereby identifying equations that produce the most accurate out-of-sample point predictions. By using P_j^2 additional equations are often identified for further consideration in model building. Observe from (3) that the PRESS residuals in P_j^2 are a function of both least squares residuals e_i and leverage values h_{ii}. Thus, the influence of both extreme Y_i and X_i values are reflected in this statistic. The global maximum for P_j^2 may not necessary be the same equation or even in the same subset as the other variable selection criteria. When the global maximum for P_j^2 differs markedly from other criteria (as in Table 1), one should investigate the possibility that the model is underspecified. The PRESS statistic and/or leverage values are available from most statistical software packages. Hence, P_j^2 is easily calculated from (10). Given the behavior and merits of this statistic, its use in statistical modeling has become increasingly popular. Indeed, P_j^2 adds another dimension of prediction accuracy to the analysis.

REFERENCES

Akaika, H. (1969), "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21,243-247.

Allen, D. M. (1984), "Discussion," *Technometrics*, 26, 319-320.

Amemiya, T. (1980), "Selection of Regressors," *International Economic Review*, 21,

331-354.

Belsley, D.A., Kuh E., and Welsch R. E.. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, NY: Wiley.

Feldstein, M. S. (1973), "Multicollinearity and the mean square error of Alternative Extimators," *Econometrica,* Vol. 41.

Hoaglin, D.C., and Welsch, R.E. (1978), 'The HAT Matrix in Regression and ANOVA', *The American Statistician* 32, 17-22.

Hocking, R. R. (1976), "The analysis and selection of variables in multiple regression," *Biometrics*, 1-49.

Hocking, R. (2003), *Methods and Applications of Linear Models*, 2nd ed., New York: John Wiley.

Maddala, G.S. (2001), *Introduction To Econometrics*, 3rd., NY, John Wiley.

Mallows, C.L. (1973), "Some Comments on C_p," *Technometric,* 15, 661-75.

Montgomery, D.C., Peck, E.A. and Vining (1992), *Introduction to Linear Regression Analysis*, 3rd. ed., New York: John Wiley.

Myers, R.H. (1990), *Classical and Modern Regression With Application, Boston*, 2nd edition, Mass, Duxbury.

Navy Manpower and Material Center (1979), "Procedures and Analysis for Staffing Standards: Regression Analysis Handbook," San Diego, CA.

Walls, R.E., and D.L. Weeks (1969), "A Note on the Variance of a Predicted Response in Regression," *The American Statistician*, 23, 24-26.