A Comparison of the Coefficient of Predictive Power, the Coefficient of Determination and *AIC* for Linear Regression

Francis A. Méndez Mediavilla
Texas State University

Frank Landram
West Texas A&M University

Vivek Shah
Texas State University

This paper describes the development and relevance of the coefficient of prediction, P^2. The coefficient of prediction is independent from the quality of fit. This paper shows evidence of the gain in accuracy and precision in the estimate of the quality of prediction when using P^2.

INTRODUCTION

The coefficient of determination, R^2, is widely used as a measure of fit for model selection and variable selection. However, R^2 does not offer any insight about the quality of the predicted values or the potential influence of particular observed values on the predicted values. In regression analysis, statistical procedures such as *t* test, F test, and prediction intervals provide insight into the quality and the potential influence of individual observations on the estimate.
The prediction sum of squares (PRESS statistic) is used as an indication of the predictive power of a model. Computing the PRESS statistic consists on fitting the model, repeatedly, leaving out an observation each time. In each repetition the model is used to predict the observation that was left out.

PRESS simulates prediction by leaving out the observation that it is trying to predict. An *external residual* for the i^{th} observation is equivalent to calculating the external predicted value $\hat{Y}_{(i)}$ without the use of the i^{th} observation. Since Y_i is not used in fitting the regression model, both the external predicted values and the external residuals are independent of Y_i.

The PRESS statistic is the sum of the squared external residuals (see FIGURE 1).

FIGURE 1
PREDICTION SUM OF SQUARES

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2$$

where,

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

The PRESS is a measure of how well the use of the fitted values for a subset model can predict the observed responses, Y_i (Neter *et al.*, 1996). The *best* regression will have a relatively small predictive sum of squares. It has been shown that PRESS is a weighted function of the least squares residuals. Residuals associated with observations whose prediction variability is large are weighted less (Quan, 1988). Observe the distinction made between external predicted values $\hat{Y}_{(i)}$, which are independent of Y_i, and internal predicted values $\hat{Y}_i$. The independence of the external residuals in
FIGURE 1 enables the PRESS statistic to be a true assessment of the validity or prediction capabilities of the regression model.

Internal predicted values come from a single regression model that includes all n observations in its construction. Thus, unlike external predicted values $\hat{Y}_{(i)}$, internal predicted values $\hat{Y}_i$ are not independent of Y_i. Internal or ordinary residuals,

FIGURE 2
INTERNAL OR ORDINARY RESIDUALS

$$e_i = Y_i - \hat{Y}_i$$

The sum of squares error (SSE) is the sum of the squared internal residuals (see FIGURE 3).

FIGURE 3
SUM OF SQUARE RESIDUALS

$$SSE = \sum e_i^2$$

While SSE measures quality of fit, PRESS measures quality of prediction. Surprisingly, the computations involved computing n external residuals and the PRESS statistic are minimal. The computations and methodology involved in calculating external residuals, the PRESS statistic, and the P^2 are described below. The performance of P^2 is compared to the performance of R^2 and the Akaike Information Criterion (*AIC*) on a subsequent section

MEASURES OF PREDICTION

Consider the following regression equation: $Y_i = \sum_{j=0}^{k} b_j x_{ij} + e_i$

where $x_{i0} = 1$ for all i and $i = 1,2,...,n$. Using matrix-vector notation for convenience, the appropriate normal equations for this regression are where $\mathbf{X}$ is an $n \times p$ data matrix, $\mathbf{Y}$ is an $n \times 1$ column vector for the response variable and $\mathbf{b}$ is $p \times 1$ vector of estimated regression coefficients $(p = k + 1)$.

FIGURE 4
NORMAL EQUATIONS

$$\mathbf{X'Xb} = \mathbf{X'Y}$$

Solving (FIGURE 4) for $\mathbf{b}$ consists of obtaining the inverse to the $p \times p$ matrix $(\mathbf{X'X})$ and multiplying it by the $p \times 1$ column vector $\mathbf{X'Y}$.

FIGURE 5
VECTOR OF ESTIMATED REGRESSION COEFFICIENTS

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

The *HAT* Matrix

The *Hat matrix* is an extremely efficient matrix for calculating prediction measures as well as for internal data analysis (Hoaglin, et.al., 1973, Hemmerle, 1967). This $n \times n$ matrix is defined in FIGURE 6.

FIGURE 6
***HAT* MATRIX**

$$\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$$

where the data and inverse matrices were defined earlier. Properties of the Hat matrix are (a) $0 \le h_{ii} \le 1$ and (b) $\sum h_{ii} = p$.

(a) The diagonal elements of $\mathbf{H}$ are between zero and one.

(b) Unless $(\mathbf{X'X})^{-1}$ is ill-conditioned, the diagonal elements of $\mathbf{H}$ sum to p.

Using the Hat matrix, a $n \times 1$ vector of point predictions is easily computed as shown in FIGURE 7.

FIGURE 7
VECTOR OF POINT PREDICTORS

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

Estimated prediction variances can be obtained as shown in FIGURE 8.

FIGURE 8
ESTIMATED PREDICTION VARIANCES

$$\hat{\mathbf{V}}(\hat{\mathbf{Y}}_{(i)}) = (\mathbf{MSE}_p)h_{ii}$$

where h_{ii} the ith diagonal element of $\mathbf{H}$ and $\mathbf{MSE}_p$ is the mean square error for the p parameter model. A $n \times 1$ vector of internal residuals may be computed as follows,

FIGURE 9
VECTOR OF INTERNAL RESIDUALS

$$\mathbf{e} = [\mathbf{I} - \mathbf{H}]\mathbf{Y}$$

External Residuals and *PRESS*

Computationally, one of the most useful properties the Hat matrix its use computing external residuals (FIGURE 10) and the PRESS statistic (FIGURE 1).

FIGURE 10
EXTERNAL RESIDUALS

$$e_{(i)} = \frac{e_i}{(1 - h_{ii})}$$

where e_i and h_{ii} are defined above. The external residual for the i^th^ observation, $e_{(i)}$, is obtained from computing $\hat{Y}_i$ in FIGURE 7without the use of the i^th^ observation Y_i. Since Y_i was not used in computing the regression model to predict $\hat{Y}_i$, the external residual $e_{(i)}$ is independent of Y_i. This enables the PRESS statistic to be a true assessment of the prediction capabilities of the regression model. In short, external residuals enable quality of prediction measures to be independent of quality of fit measures (i.e., R^2).

The *AIC* Statistic

Of the several ways different ways of trading off goodness-of-fit and parsimony, adjusted R^2 has the least amount of adjustment for extra explanatory variables. The most popular alternatives are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC or Schwartz Information Criterion) (Kennedy, 1998).

FIGURE 11
INFORMATION MEASURES (AKAIKE & BAYESIAN)

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2p$$

$$BIC = n \ln\left(\frac{SSE}{n}\right) + p \ln(n)$$

These *information measures* statistics are designed to select, from a small number of alternatives, the model with the best predictive power. Usually, the model that gives the smallest value of AIC (or BIC) statistic is the preferred one. The BIC is claimed to be an improvement over the AIC since the AIC is inclined to overfitting the data. From the equations inFIGURE 11, it is also evident that AIC and BIC are not independent from Y_i.

The P^2 Statistic

Observe that the PRESS statistic in FIGURE 1 is similar to the sum of square errors in regression analysis in FIGURE 12. The PRESS statistic uses external predicted values $\hat{Y}_{(i)}$ while the SSE statistic uses internal predicted values $\hat{Y}_i$. Thus, just as SSE is used in calculating the coefficient of determination,

FIGURE 12
SUM OF SQUARE ERRORS

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

FIGURE 13
COEFFICIENT OF DETERMINATION

$$R^2 = 1 - \frac{SSE}{SST}$$

where,

$$SST = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

The PRESS statistic is used in calculating P^2 as shown in FIGURE 14.

FIGURE 14
COEFFICIENT OF PREDICTIVE POWER

$$P^2 = 1 - \frac{PRESS}{SST_{(i)}}$$

where,

$$SST_{(i)} = \sum\left(Y_i - \overline{Y}_{(i)}\right)^2 = \left[\frac{n}{n-1}\right]^2 SST$$

The symbol $\overline{Y}_{(i)}$ represents the arithmetic mean for $n-1$ values of Y_i; the value of Y_i is subtracted from the mean $\overline{Y}$ used in SST. Thus, unlike Y_i and $\overline{Y}$ in FIGURE 13, Y_i and $\overline{Y}_{(i)}$ in FIGURE 14 are independent (see proof in appendix). This independence concept is the same concept associated with the PRESS statistic. Notice that all components in IGURE 14 are totally independent. This independence enables P^2 to adequately measure a regression model's quality of prediction.

Notice that, in computing the coefficient of prediction, P^2, the ratio PRESS to $SST_{(i)}$ is used, rather than the ratio SSE to SST (see equations FIGURE 13 and FIGURE 14). Thus, P^2 is analogous to R^2; however, P^2 measures quality of prediction while R^2 measures quality of fit. The use of P^2 as a model selection criterion, addresses the loss in predictability when a variable is deleted from the regression model.

A comparison between P^2, R^2 and AIC is shown in the following sections. A Monte Carlo simulation is used to evaluate the efficiency of P^2, R^2 and AIC. A real data set is used to demonstrate the performance of these measures.

Monte Carlo Experiment: Comparing P^2, R^2 and AIC

A Monte Carlo experiment was performed to compare the performance of the P^2, against the coefficient of determination and the AIC. Given a data matrix $\mathbf{X} \sim Gaussian(0,\Sigma)$, linear additive models of the forms: $x_3 = f(x_1)$, $x_3 = f(x_2)$ and $x_3 = f(x_1, x_2)$ were fitted. Samples sizes 1,000 are extracted from the population $\mathbf{X}$. Linear models are fitted to these sample data.

Estimates for the sampling distribution for P^2, the coefficients of determination, the AIC and the BIC are obtained. Several covariance matrices were used to specify various types of relationships between variables. The linear models shown in FIGURE 15 are fitted at each iteration.

FIGURE 15
REGRESSION MODELS FOR THE MC SIMULATION

$$X_3 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + e$$
$$X_3 = \hat{\beta}_0 + \hat{\beta}_2 X_2 + e$$
$$X_3 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e$$
$$X_3 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 + e$$

One thousand iterations were run. Each iteration consists of the following steps: 1) sample 1,000 observations from a multivariate Gaussian distribution with the given covariance matrix; 2) generating a sample size 1,000; 3) estimating the parameters for the model; 4) computing R^2 and P^2 for the training set; 5) using the resulting model to estimate X_3 for the hold-out set; 6) estimate the sampling distributions for R^2, AIC and P^2; 7) compare the relative efficiency of P^2.

The first covariance matrix Σ_1 (see FIGURE 16) was specified so that one predictor (x_1) is strongly related to the response variable (x_3), while x_2 is almost not related to the response variable and not related at all to the other predictor.

FIGURE 16
COVARIANCE MATRIX

$$\sum_1 = \begin{bmatrix} 1 & 0 & 0.90 \\ 0 & 1 & 0.10 \\ 0.90 & 0.10 & 1 \end{bmatrix}$$

The 95% percentile estimator $\left[F_\theta^{-1}(0.025), F_\theta^{-1}(0.975)\right]$ is computed for the sampling distributions for the statistics of interest.

TABLE 1
MODEL - $x_3 = \hat{\beta}_0 + \hat{\beta}_1 x_2$

Statistic	$F_\theta^{-1}(0.025)$	$F_\theta^{-1}(0.975)$
P^2	-0.0037	0.0320
R^2	0.0003	0.0358
AIC	1351.662	1475.128
BIC	1364.306	1487.772

Evidently, adding to the model variable x_1 as a predictor to our model as a predictor will significantly increase the model fit.

TABLE 2

MODEL - $x_3 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Statistic	$F_\theta^{-1}(0.025)$	$F_\theta^{-1}(0.975)$
P^2	0.7871	0.8466
R^2	0.7887	0.8479
R^2_{adj}	0.7879	0.8473
AIC	506.3423	627.3765
BIC	523.2007	644.2349

Adding the interaction term to the model does not improve the fit by much.

TABLE 3

MODEL - $x_3 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$

Statistic	$F_\theta^{-1}(0.025)$	$F_\theta^{-1}(0.975)$
P^2	0.7872	0.8462
R^2	0.7898	0.8481
R^2_{adj}	0.7886	0.8471
AIC	506.8535	627.8151
BIC	527.9265	648.8882

We are defining the relative efficiency as: $\frac{CV_\theta}{CV_{P^2}}$; where CV_{P^2} is the coefficient of variability for P^2 and CV_θ is the coefficient of variability for statistic θ.

TABLE 4

RELATIVE EFFICIENCY, $x_3 = \hat{\beta}_0 + \hat{\beta}_1 x_2$

Statistic θ	Coefficient of Variation	Relative Efficiency P^2
P^2	1.2192	1.0000
R^2	0.8002	0.6563
R^2_{adj}	0.9639	0.7906
AIC	0.0218	0.0179
BIC	0.0216	0.0177

TABLE 4 shows that the relative efficiency of P^2 is poor for this model with relatively poor fit. However, in repeated simulations it was found that for well-fitted models the P^2 tends to be relatively more efficient than the AIC, e.g. TABLE 5

TABLE 5
RELATIVE EFFICIENCY, $x_3 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Statistic θ	Coefficient of Variation	Relative Efficiency P^2
P^2	0.0183	1.0000
R^2	0.0181	0.9891
R^2_{adj}	0.0182	0.9945
AIC	0.0551	3.0109
BIC	0.0531	2.9016

Several Monte Carlo simulations of a similar kind were conducted. The results were similar as the ones shown previously. The P^2 statistic is not bound by zero and one, as is the coefficient of determination, R^2. However, in a practical sense, P^2 rarely exceeds the value of one. In several simulation experiments of over 1000 data sets, $P^2 \leq 1.0$ was true in every case and $P^2 < 0.0$ in some cases.

A Practical Example: Comparing P^2, R^2 and AIC

Consider the data of vehicles from 2005. There is data from 411 vehicles. The mileage per gallon in the city will be estimated based on the horsepower and the inverse of the weight of the vehicle.

Y: Mileage per gallon in city

X_1: Horsepower

X_2: Weight $^{-1}$

A linear model is suggested in FIGURE 17.

FIGURE 17
LINEAR MODEL

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e$$

TABLE 6
CORRELATION MATRIX

	city.mpg	horsepower	1/weight
city.mpg	1	-0.7317	0.8899
horsepower	-0.7317	1	-0.6500
1/weight	0.8899	-0.6500	1

The regression estimates are shown in TABLE 7.

TABLE 7
LINEAR REGRESSION

	Estimate	pvalue
Intercept	8.304	0.0000
Horsepower	-0.0151	0.0000
Weight $^{-1}$	0.0005	0.0000
R^2	0.8327	
F test (pvalue)	0.0000	

The performance of R^2, P^2 and *AIC* will be compared. The data set consists of 411 observations. Bootstrap estimates for R^2, P^2 and *AIC* and their corresponding biases and standard errors were estimated using 500 replicas.

The results are shown in

TABLE **8**. Let $\hat{\theta}^*$ be the bootstrap estimate, SE($\hat{\theta}$) denote the standard error of the estimate $\hat{\theta}$, the relative bias is defined as $bias(\hat{\theta})/\hat{\theta}^*$ and the relative standard error is defined as $se(\hat{\theta})/\hat{\theta}^*$.

TABLE 8
BOOTSTRAP ESTIMATES

θ	$\hat{\theta}^*$	Bias ($\hat{\theta}$)	SE ($\hat{\theta}$)	Relative Bias	Relative SE($\hat{\theta}$)
R^2	0.8327	-1.6690	0.1270	-2.0044	0.1526
AIC	1589.415	8.5406	36.6790	-0.0054	0.0231
P^2	0.8303	0.0034	0.0143	0.0042	0.0172

It was found that, for this *real life* data set, the P^2 is relatively more accurate and more precise than the coefficient of determination and the *AIC*.

FURTHER ANALYSIS

The simulation in this study was limited to a few variables exhibiting specific multivariate Gaussian distribution. Further analysis would involve models with a larger number of variables, varying the covariance structure at random and *benchmark* data sets.

The usefulness of P^2 should be tested in an experiment involving model selection. The consistency of P^2 must also be evaluated.

CONCLUDING REMARKS

In regression analysis, the PRESS concept is used in generating quality-of-prediction measures. Since PRESS is independent of Y_i, it is a true assessment of the prediction capabilities of the regression model. A relative measure of prediction is P^2, the coefficient of prediction. Unlike R^2 and *AIC*, P^2 is based on PRESS and thus is also independent of Y_i. The numerical limits of P^2 are not constrained to values between zero and one. Although theoretically possible, P^2 rarely exceeds the value of one. From Monte Carlo simulations it was found that P^2 can be more efficient than *AIC* for the same sample sizes. From a real life example it was found that P^2 was more accurate and more precise than both: R^2 and *AIC*. The implications would be that using P^2 for model selection would represent a lesser probability of type II error than the classic test for increments in R^2 or tests for reductions in AIC.

APPENDIX

The concept of dividing the PRESS statistic by $SST_{(i)}$ rather than SST is unique with this author. Calculating P^2 by dividing by $SST_{(i)}$ gives a more accurate measure of quality of fit.

$$P^2 = 1 - \frac{PRESS}{SST_{(i)}}; \tag{1}$$

where

$$SST_{(i)} = \sum \left(Y_i - \overline{Y}_{(i)}\right)^2 = \left[\frac{n}{n-1}\right]^2 SST. \tag{2}$$

$$SST = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \tag{3}$$

The symbol $\overline{Y}_{(i)}$ represents the arithmetic mean of $n-1$ Y-values; the current Y_i values is subtracted out of $\overline{Y}$. Thus,

$$Y_{(i)} = \frac{\left(n\overline{Y} - Y_i\right)}{(n-1)}; \tag{4}$$

where $\overline{Y}$ is the overall arithmetic mean for n Y-values. Using (30) for $\overline{Y}_{(i)}$ we show

$$\begin{aligned} SST_{(i)} &= \sum \left(Y_i - \overline{Y}_{(i)}\right)^2 \\ &= \sum \left[\frac{\left(Y_i(n-1) - \left(n\overline{Y} - Y_i\right)\right)}{(n-1)}\right]^2 \\ &= \sum \left[\frac{\left((n-1+1)Y_i - n\overline{Y}\right)}{(n-1)}\right]^2 \\ &= \sum \left[\frac{\left(n\left(Y_i - \overline{Y}\right)\right)}{(n-1)}\right]^2 \\ &= \left[\frac{n}{(n-1)}\right]^2 \sum \left(Y - \overline{Y}\right)^2 \\ &= \left[\frac{n}{(n-1)}\right]^2 SST \end{aligned} \tag{5}$$

Hat matrix

$$
\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{Xb} \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{Hy}
\end{aligned}
$$

where,

$$\mathbf{H} = \{h_{ij}\} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and

$$\mathbf{H} = \mathbf{H}'\mathbf{H} \qquad (6)$$

REFERENCES

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. Technometrics, *13*(3), 469-475.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley.

Berenson, M. L., Levine, D. M., & Goldstein, M. (1983). Intermediate statistical methods and applications. Englewood Cliffs, N.J.: Prentice-Hall.

Cook, R. D. (1977). Detection of influential observations in linear regression. Technometrics, *19*, 15-18.

Cook, R. D. (1979). Influential observations in linear regression. Journal of the American Statistical Association, *74*, 169-174.

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman & Hall.

Efron, B., & Tibshirani, R. J. (1998). An Introduction to the Bootstrap. New York: Chapman & Hall/CRC.

Goodnight, J. H. (1979). A tutorial on the sweep operator. The American Statistician, *33*, 149-158.

Graybill, F. A. (1976). Theory and application of the linear model. Boston, Massachusetts: Duxbury.

Hemmerle, W. J. (1967). Statistical computations on a digital computer. Waltham, Massachusetts: Blaisdell.

Hoaglin, D. C., & Welsch, R. E. (1973). The hat matrix in regression and anova. The American Statistician, *32*, 17-22.

Kennedy, P. (1998). A Guide to Econometrics. Cambridge, Massachusetts: The MIT Press.

Mallows, C. L. (1973). Some comments on cp. Technometrics, *15*, 661-675.

Myres, R. H. (1986). Classical and modern regression with application. Boston, Massachusetts: Duxbury.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models (4th ed.). Chicago: Irwin.

Pao, P. (1971). Some notes on misspecification in multiple regression. The American Statistician, *25*, 37-39.

Quan, N. T. (1988). The prediction sum of squares as a general measure for regression diagnostics. Journal of Business and Economic Statistics, *6*(4), 501-504.

Schatzoff, M., Tsao, R., & Fienberg, S. (1968). The efficient calculation of all possible regression. Technometrics, *10*, 769-779.

Tarpey, T. (2000). A note on the prediction sum of squares statistic for restricted least squares. The American Statistician, *54*(2), 116-118.