Example for Paper MS Word Template for BOMSR: Top Margin: Left=1.5”, Font: size= 16 pt., Font Face =Arial & Bold, Align= Center, Line Space=single

Formatted: Right: -0.38", Space Before: 0 pt, After: 0 pt

Medikondu Kishore[1*] Janardhan.M.
Départment of Mathematics, SVR College of Engineering and Technology
Andhra Pradesh, India

Formatted: Centered

ABSTRACT

Research papers should be accompanied by an abstract, which will appear in front of the main body of the text. It should be written in complete sentences and should summarize the aims, methods, results and conclusions in less than 250 words. The abstract should be comprehensible to readers before they read the paper and abbreviations, citations and mathematical equations/notations should be avoided.

Margin: Left=1.5 inch, Right=1.5 inch, Font: size=9 pt. Font Face =Arial & italic Align= justify Line Space=single

Keywords: Regression, data structure, prediction, simulation.

1. INTRODUCTION

Paragraph: Margin: Left=1 inch, Right=1 inch, Font: size10 pt. Font Face =Arial, Align= justify, Line Space= 1.3 pt.

Formatted: Font: 10 pt, Not Italic

Formatted: Indent: Left: 0", Right: 0", Line spacing: Multiple 1.3 li

Formatted: Font: 10 pt, Not Italic

Formatted: Font: 10 pt, Not Italic

In the establishment of the prediction model, three stages are fundamental: possible selection of the variables, the estimation of the coefficients of the variables selected and the validation of the model. Ideally, this validation should be done on different observations. But in most practical situations, the selection of the variables, the estimation of the coefficients and the validation are done using the same sample. Indeed, it is often difficult to have separate samples for the various stages of modeling, because the dataset available to the researcher is frequently too small to use part of it to establish the regression model and the remaining for its validation. Sometimes, the number of predictors is higher than the number of observations.

The objective of this work is to bring some useful information for the users, especially those who do not have the possibility to validate the models from external data. In a more concrete way, we propose to examine the predictive value of a regression model by calculating a coefficient, similar to the multiple coefficient of determination, which we call coefficient of determination of prediction. It is denoted R_p^2 and is defined, for n_p new observations, as follows:

$$R_p^2 = 1 - \sum_{i=1}^{n_p} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n_p} (y_i - \bar{y})^2 .$$

In this relation, y_i indicates the actual value of the dependent variable for the new individual i ($i=1,\ldots,n_p$). $\hat{y}_i$, is the predicted value for this individual given by the regression model, $\bar{y}$ is the arithmetic mean of n observations of the dependent variable in the sample which was used to establish the model.

2. GENERATION OF THE DATA

The realization of this work supposes the availability of a great number of repetitions of samples responding to the same known theoretical model. In practice, as the theoretical model is unknown, we use the Monte-Carlo method based on the generation of the data by computer according to a fixed theoretical model.

2.1. Theoretical model

We consider the traditional theoretical model of multiple linear regressions as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\ ,$$

where $\mathbf{y}$ is an $n\times 1$ vector observations of the dependent variables, $\mathbf{X}$ is the matrix $n\times k$ of k explanatory variable, $\boldsymbol{\varepsilon}$ the vector of n theoretical residuals and $\boldsymbol{\beta}$ the vector of the theoretical regression coefficients. It is supposed that the residuals are independent random variable of the same normal distribution of null mean and constant variance σ^2. The parameters to be simulated are $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, while the vector $\mathbf{y}$ is calculated by the model.

2.2. Controlled factors

The factors controlled for the theoretical models are the number of explanatory variables (k), the number of observations (n), the index of collinearity of the explanatory variables (IC), the index of decrease of the regression coefficients (Ib) and the theoretical coefficient of determination (R_0^2).

$$\beta_i = c(Ib)^{i-1} \qquad (i = 1,\ 2,\ 3,\ 4,\ 5)$$

where β_i is the value of coefficient i, Ib the index of decrease of the regression coefficients and c a constant.

2.3. Methods of regression studied

On the one hand, we considered the classical method of least squares without variables selection and on the other hand, the *stepwise* selection method of variables is used. These methods were adopted, because they are among the most used methods, and are available in almost all statistical software.

The selection of variables is based on the *t* test of Student or *F* test of Snedecor for significance of the regression coefficients. We used the same level of significance for the introduction and the exclusion of a variable in the model. Two theoretical levels were retained: 0.15 and 0.05.

3. RESULTS

3.1. Effects of the various factors on the coefficient R_p^2

The analysis of table 1 shows that coefficient R_p^2 is more often lower than the theoretical coefficient of determination. The ratio increases as the sample size increases, for a given value of k and R_0^2.

Table 1: Average observed values of R_p^2, expressed in proportion of R_0^2, according to k, n and R_0^2.

k	n	$R_0^2 = 0.20$	$R_0^2 = 0.40$	$R_0^2 = 0.60$	$R_0^2 = 0.80$
Complete model					
5	8	-14.39	-4.76	-1.54	0.06
10	200	0.82	0.93	0.97	0.99
30	50	-0.15	0.52	0.77	0.90
30	600	0.91	0.96	0.98	0.99
Model selected ($\alpha = 0.15$)					
5	8	-1.66	-0.31	0.26	0.65
5	100	0.79	0.92	0.96	0.98
10	17	-0.99	0.15	0.56	0.81
10	200	0.89	0.96	0.98	0.99
30	50	-0.40	0.44	0.75	0.90
30	600	0.93	0.98	1.00	1.00
Model selected ($\alpha = 0.05$)					
30	600	0.93	1.00	1.00	1.00

For known values of R_0^2 and k/n, the ratio R_p^2/R_0^2 depends little on the values of *k* and *n*. We also note that the ratio is weaker for the low values of R_0^2. Finally, the use of variables selection tends to increase the ratio.

3.2. Determination of the levels of factors combinations leading to a null predictive value

In order to obtain results easily usable in practice, we determined the validity limits of the equations for the purpose of prediction by being unaware of the effect of factors IC and Ib on the prediction. These limits are obtained by determining the levels of the ratio k/n leading to a zero value of R_p^2. These levels give on average the thresholds of combinations of factors from which the model led to predictions of quality lower than the prediction given by the arithmetic mean of the dependent variable of the sample.

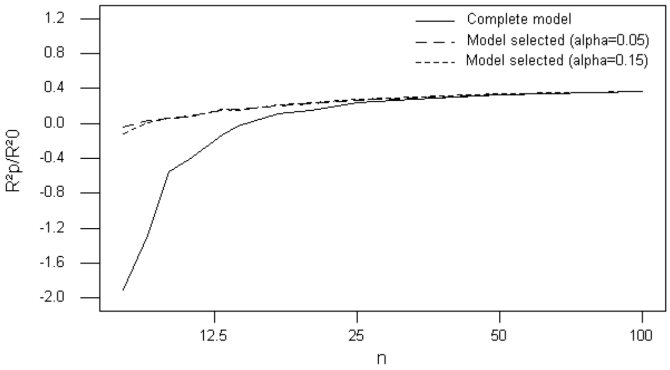

Figure 1. Evolution of the ratio R_p^2 / R_0^2 according to the sample size on logarithmic scale in X-coordinate, for $k = 5$, R_0^2 =0.40.

From this table, we note that this size varies according to the method used to establish the model. It is higher for the complete models and decreases gradually with the intensity of the selection. It also decreases as the theoretical value R_0^2 increases.

4. DISCUSSION AND CONCLUSION

Several authors documented criteria that assess the quality of a model. These criteria are based on the difference between the estimated model and the presumed known theoretical model. In the present study, the criterion used compares to new observations resulting from the same population as individuals of the sample, the variability of the errors of prediction, when the predictions are carried out by a regression equation and on the other hand when these predictions are equal to the arithmetic mean $\bar{y}$ of the dependent variable in the sample. It thus gives an idea of the improvement of the quality of prediction by taking into account the explanatory variables. It also informs about the validity limits of a prediction model.

The plan of simulation considers data of varied structures. In particular, we considered the case where all the explanatory variables available are indeed present in the theoretical model ($k \leq 5$) and the case where certain explanatory variables available are not present in the theoretical model. This approach makes it possible to be close to the situations often encountered in practice.

5. REFERENCES

Paragraph: Margin: Left=1 inch, Right=1 inch, Font: size10 pt. Font Face =Arial, Align= justify, Line Space= single pt.

Akossou, A.Y.J., 2005, *Impact de la structure des données sur les prédictions en régression linéaire multiple*. PhD Thesis, Fac. Univ. Sci. Agron., Gembloux, Belgium, 215 p.

Bendel, R.B., Afifi, A.A., 1977, Comparison of stopping rules in forward stepwise regression. *J. Amer. Stat. Assoc.* **72**, 46-53.

Copas, R.D., 1983, Regression, prediction and shrinkage. *J. R. Stat. Soc.* **B 45**,311-354.

Dempster, A.P., Schatzoff, M., Wermuth N., 1977, A simulation study of alternatives to ordinary least squares. *J. Amer. Stat. Assoc.* **72**, 77-106.

Meg, B. C., 1988, Determining the optimum number of predictors for linear prediction equation. *Amer. Meteo. Soc.* **116**, 1623-1640.

Miller, A.J., 1990, *Subset selection in regression*. Monographs on statistics and applied probability 40. Chapman and Hall.

Palm, R., De Bast, A., Lahlou, M., 1991, Comparaison des modèles agrométéorologiques de type statistique empirique construits à partir de différents ensembles de variables météorologiques. *Bull. Rech. Agron.* Gembloux **26**, 71-89.

Roecker, E.B., 1991, Prediction error and its estimation for subset-selected models. *Technometrics* **33**, 459-468.