Using Stepwise Method to Find the Most Influencing Feature to the Cell Nuclei of a Breast Mass

Pengwei Huang

San Diego State University, San Diego, CA 91208, USA

Phuang4@sdsu.edu

Abstract. In real life, the influencing factors that lead to an outcome are varied and complex. However, how to organize the complex variable relationship into a concise and effective model for prediction is critical. Therefore, the Stepwise method was used in this study to organize the relationship between malignant and benign detection of breast cancer tumors and changes in cell characteristics into a model. In this paper, the diagnosis data of breast cancer of patients in the hospital were quoted. The benign and malignant characteristics of breast cancer were taken as independent variables, and the characteristic changes of cells were taken as dependent variables to conduct data analysis with R language, so as to obtain the most effective model. The results show that the prediction accuracy of the model obtained by AIC method is the highest. AIC selected seven dependent variables and reached 79.8 percent. The model prediction accuracy of BIC was 77.2 percent. Compared with AIC method, BIC only selected four dependent variables and obtained a more concise model. However, BIC is too concise and loses the accuracy of model prediction in some aspects.

Keywords: Breast Cancer, Stepwise Method, Model Selection, Cell Nuclei Feature.

1. Introduction

In daily life, people will encounter all kinds of phenomena. Like the movement of clouds, the rise and fall of ocean waves and the migration of animals. However, observing these phenomena is only superficial, and it is the most critical to analyze the causes and make predictions based on these phenomena. Just as wind speed affects the movement and direction of clouds, the factors that affect how things develop can be complex and varied. The purpose of this study is to effectively select the influence of certain factors on the results. Usually, the factors that affect the outcome are called independent variables and the outcome is called dependent variables. A dependent variable may be affected by dozens or even hundreds of independent variables. However, the influence of some independent variables on the dependent variable is not so important.

If all the independent variables that may affect the dependent variable are selected, the model will become extremely complicated and inefficient. Therefore, some scholars have made researches on this and obtained some criterion and methods for selecting variables. In the book "Empirical Model-Building and Response Surfaces" co-authored with Norman R. Draper, George Edward Pelham Box, creator of the Box-Jenkins model (the ARIMA model) and the Box-Cox transformation, wrote: “essentially, all models are wrong, but some are useful.” [1] A thought-provoking words tell people: statistical inference model used is determined based on the incomplete information, and the incomplete information makes mistakes is inevitable, but this does not mean statistical modeling is pointless tricks, on the contrary, as long as we can to control the error within a certain range, then this model is useful for practical application.

In fact, incomplete information is the premise and foundation of statistical inference. Although incomplete information makes statistical inference have its significance and value, at the same time, it also makes it possible for different people or even the same person to establish different models when facing the same problem. Therefore, a realistic and important problem is placed in front of people, it is the model selection. Model selection seems to be expected to perform the function of selecting, among all possible models, the one that most closely approximates the real situation, or in other words, the one with the least error or loss. Unfortunately, this is almost impossible. Here's the

problem: People can't exhaust all possible models and don't know what's really going on, they don't have a single measure of error or loss. Therefore, the question of model selection, like the question of model building, is a matter of different opinions.

At the end, people expect to choose a model that makes the best predictions for unknown data. However, people have no way to accurately calculate effect of a model to predict unknown data, only through a certain method to estimate a model of unknown data prediction effect, at the same time, what is called prediction well, does not have a unified standard, it depends on the actual problem, the actual demand, the angle of view, and so on. Therefore, the combination of different methods to estimate the actual prediction effect of the model and different criteria to measure the prediction effect of the model constitute a variety of model selection methods. The estimation methods of the actual prediction effect of the model can be summarized as the following three categories:

1) The prediction effect of the model in the sample is used to estimate the actual prediction effect of the model, such as the mean square error of the forecast in the sample

2) The actual prediction effect of the model is estimated by the weighted sum of the prediction effect of the model in the sample and the complexity of the model, such as AIC and BIC

3) The out-of-sample prediction effect of the model is used to estimate the actual prediction effect of the model: for example, cross validation

The research topic of this paper is how to select the most effective and accurate model in complex data. Therefore, this paper adopts the second method, using AIC and BIC as research methods. This paper will use AIC and BIC to get the model, compare the accuracy rate to select the optimal model, and analyze the advantages and disadvantages of the two methods in selecting the model.

2. Methodology

In the Breast Cancer Wisconsin (Diagnostic) Data Set, 569 patients were reported. The diagnosis results are divided into benign and malignant, in which B represents benign and M represents malignant. Of these, 357 patients were diagnosed as benign and 212 were diagnosed as malignant. The data also include ten real -valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these 10 features are computed respectively. So, this data includes a total of 30 variables for nuclear characteristics [2]. These 30 features will be used as independent variables in the study of model selection and prediction of the dependent variable diagnosis results. 30 independent variables is not very complicated compared to a model, but it is still not efficient and concise. So, some of the variables will be removed in the model selection to make the model less complex. In the following ,this article will show in detail how to simplify this model.

First of all, divide this data set into two matric Y and X. Set Y matric as response variable diagnosis (“1” represents M, “0” represents B) and set X matric as all of the independent variables (column 3:32 of data set). Then build a data frame “ data1” by combining matric Y and X. Since there are 30 independent variables, severe multicollinearity could exist in this data set. In this paper, VIF method is used to calculate the multicollinearity of each parameter. VIF (variance inflation factor) is the ratio of the variance of estimating some parameter in a model that includes multiple other parameters by

the variance of a model constructed using only one term [3]. In this data, the regression model can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + \epsilon. k = 30 \quad (1)$$

Calculating VIF factor for each $\widehat{\beta_i}$ can get the multicollinearity between the various parameters. When calculating VIF factor, it is necessary to use the coefficient of determination of the variables. The VIF factor formula is:

$$VIF_i = \frac{1}{1-R_i^2} \quad (2)$$

According to this formula, VIF values of 30 independent variables are calculated as following:

Table 1. VIF value of each parameters

smoothness_se	texture_mean	area_se
4.027923	11.884048	41.163091
smoothness_mean	texture_worst	concavity_mean
8.194282	18.569966	70.767720
concave.points_se	compactness_worst	perimeter_worst
11.520796	36.982755	405.023336
fractal_dimension_mean	perimeter_se	symmetry_se
15.756977	70.359695	5.175426
concave.points_worst	area_mean	smoothness_worst
36.763714	347.878657	10.923061
concave.points_mean	radius_mean	concavity_se
60.041733	3806.115296	15.694833
area_worst	symmetry_mean	concavity_worst
337.221924	4.220656	31.970723
perimeter_mean	fractal_dimension_se	compactness_mean
3786.400419	9.717987	50.505168
texture_se	compactness_se	radius_se
4.205423	15.366324	75.462027
symmetry_worst	fractal_dimension_worst	radius_worst
9.520570	18.861533	799.105946

There are many parameters with VIF values above 10 where indicate severe multicollinearity. Therefore, only parameters with VIF value less than 10 were selected as independent variables of the model in this study. After this screening, there are only 7 variables left. They are ***smoothness_se, texture_se, symmetry_mean, symmetry_se, smoothness_mean, symmetry_worst, fractal_dimension_se***. Combing these 7 variable with matric Y to obtain a new data set "data2". Although 7 independent variables are parsimetric enough for a model, their validity and accuracy have not been tested. Therefore, the method of Stepwise will be used to optimize a model. Before that, training data is required to optimize the model. Typically, in machine learning, 80% of the real data is used to train the model, and the remaining 20% is used to test the model. Data2 will be used for machine training and testing.

The initial model was composed of the above seven independent variables and the dependent variable Y. The initial model is called the full model. Because the research purpose of this paper is to use the Stepwise method to simplify and optimize the model, inefficient independent variables will be removed. So the stepwise direction is going to be backwards. Also, since the dependent variable takes the value of 1 or 0 (malignant or benign), the model is a logistic regression model. After building the full model, use 80% of the data2 to train the model. Here is the following result:

Table 2. Full Model Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.529	1.200	-8.774	< 2e-16 ***
smoothness_mean	44.640	11.522	3.874	0.000107 ***
symmetry_mean	-1.840	7.114	-0.259	0.795943
texture_se	1.160	0.274	4.234	2.30e-05 ***
smoothness_se	-78.297	56.502	-1.386	0.165824
symmetry_se	-123.805	22.829	-5.423	5.86e-08 ***
fractal_dimension_se	115.023	59.510	1.933	0.053256
symmetry_worst	24.509	3.776	6.490	8.57e-11 ***

The intercept and coefficients of each variable are shown in the results. For example, the coefficient of smoothness_mean is 44.640, the coefficient of symmetry_mean is -1.840, and the intercept of the model is -10.529 (the value range is three decimal places). In the result, the P value of each independent variable was also calculated. P value represents the necessary degree of independent variables in this model. The smaller P value is, the more necessary this independent variable is in the model. Conversely, the larger the P value, the more likely it is to be removed from the model. The significant level of P value is 0.05. However, Stepwise method does not only consider P value, it also needs to consider the complexity of the model. Take the AIC method that will be used next as an example.

The Akaike Information Criterion (AIC) is named after the Japanese statistician Hriotugu Akaike. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of under-fitting. AIC rewards goodness of fit, but it also penalizes increasing numbers of parameters. Because AIC discourages over-fitting of the model, the goodness of fit of the model always increases as the number of parameters increases. So the penalty for the number of parameters reduces the complexity of the model and leads to the optimal model. In model selection, a lower AIC score indicates a better model. The AIC formula is:

$$\mathrm{AIC} = -2 \times \mathcal{L} + 2 \times (\mathrm{p} + 2) \tag{3}$$

where $\mathcal{L}$ is the log-likelihood of the model, p is the number of regression parameters, excluding the intercept.

In this study, AIC made the following model selecting:

Start: AIC=460.02

Y ~ smoothness_mean + symmetry_mean + texture_se + smoothness_se + symmetry_se + fractal_dimension_se + symmetry_worst

Table 3. AIC first step score

	DF	Deviance	AIC
symmetry_mean	1	444.09	458.09
smoothness_se	1	446.00	460.00
<none>		444.02	460.02
fractal_dimension_se	1	447.41	461.41
smoothness_mean	1	460.19	474.19
texture_se	1	461.95	475.95
symmetry_se	1	476.12	490.12
symmetry_worst	1	496.05	510.05

The starting AIC score for the full model is 460.02. The starting AIC score for the initial model was 460.02. According to the AIC formula algorithm, removing either symmetry_mean or smoothness_se will optimize the model and result in a lower AIC score. Removing the

symmetry_mean parameter will result in a lower AIC score than removing smoothness_se parameter, so AIC preferentially chooses to remove this parameter.

Step: AIC=458.09

Y ~ smoothness_mean + texture_se + smoothness_se + symmetry_se + fractal_dimension_se + symmetry_worst

Table 4. AIC final score

	DF	Deviance	AIC
<none>		444.09	458.09
smoothness_se	1	446.09	458.09
fractal_dimension_se	1	447.42	459.42
smoothness_mean	1	461.69	473.69
texture_se	1	462.17	474.17
symmetry_se	1	476.70	488.70
symmetry_worst	1	513.81	525.81

In the next step, removing any of the parameters does not decrease the AIC score. So, the machine keeps all the remaining parameters according to the algorithm. The final AIC score was 458.09, and the model parameters were ***smoothness_se, texture_se, symmetry_se, smoothness_mean, symmetry_worst, fractal_dimension_se.*** After the AIC method is used, the model selection of the full model will be done again for those using BIC as a comparison.

BIC (Bayesian information criterion) is also a method of model selecting. Similar to AIC, they both use the number of parameters in the model as a penalty for model over-fitting. As with AIC, a lower BIC score indicates a better model. But BIC has a larger penalty than AIC. For this reason, the models chosen by BIC tend to be simpler than those chosen by AIC. Its formula is:

$$\mathrm{BIC} = -2 \times \mathcal{L} + \ln(\mathrm{n}) \times (\mathrm{p} + 2) \tag{4}$$

$\mathcal{L}$ is the log-likelihood of the model

p is the number of regression parameters, excluding the intercept

n is the number of data points in data set, the number of observations, or equivalently, the sample size

Compare with AIC, BIC made the following model selecting:

Start: BIC=494.77

Y ~ smoothness_mean + symmetry_mean + texture_se + smoothness_se + symmetry_se + fractal_dimension_se + symmetry_worst

Table 5. BIC first step score

	DF	Deviance	BIC
symmetry_mean	1	444.09	488.49
smoothness_se	1	446.00	490.41
fractal_dimension_se	1	447.41	491.81
<none>		444.02	494.77
smoothness_mean	1	460.19	504.59
texture_se	1	461.95	506.36
symmetry_se	1	476.12	520.53
symmetry_worst	1	496.05	540.46

At first, BIC gave the Full Model a score of 494.77. As shown in the figure, BIC, same as AIC, calculated that removing the symmetry_mean first made the model better. In contrast to AIC, BIC method calculated that removing the parameter of fractal_dimension_se could get lower BIC score. Since that, BIC remove the three parameters: symmetry_mean, smoothness_se, fractal_dimension_se and get a lower score 480.07.

Step: AIC=480.07

Y ~ smoothness_mean + texture_se + symmetry_se + symmetry_worst

Table 6. BIC final score

	DF	Deviance	AIC
<none>		448.35	480.07
texture_se	1	467.08	492.45
smoothness_mean	1	468.97	494.34
symmetry_se	1	489.70	515.07
symmetry_worst	1	539.28	564.45

According to the selection of BIC method, the final model retains the following four parameters: ***texture_se, smoothness_mean, symmetry_se, symmetry_worst***. Compared. With AIC, which retained 6 parameters, BIC only selected 4 parameters in the final model. This also proves that at most case, BIC chooses a simpler model than AIC.

According to different Stepwise methods, two different models were obtained in this study. One criterion to verify the model is the accuracy of model prediction. Next, this paper will compare the prediction accuracy of two different models on the same real data. The data that will be used for testing is the 20% of the data that was previously split, which is named test data. In logistic regression function, one of the methods to judge the binary prediction accuracy of the model is to use confusion matrix. The confusion matrix includes four parameters [4, 5]:

TP (True Positive) represents the number of benign people in this study who were correctly predicted by the model compared to the real data

FP (False Positive) represents the number of benign people in this study who were incorrectly predicted by the model compared to the real data

TN (True Negative) represents the number of malignant people in this study who were correctly predicted by the model compared to the real data

FN (False Negative) represents the number of benign people in this study who were incorrectly predicted by the model compared to the real data

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

Because the value predicted by the model is the possibility of malignancy, this paper classifies the value of prediction result greater than 0.5 as 1 (malignant), and the value of prediction result less than 0.5 as 0 (benign). After the prediction by AIC, the confusion matrix was obtained as follows:

Table 7. AIC confused matrix

	glm.pred	
Y.test	0	1
0	62	6
1	17	29

According to the confusion matrix, the number of TP was 29, the number of FP was 6 and the number of FN was 17 in the AIC model prediction results. According to the precision and recall formula, the precision of AIC model is 29 divided by 35 which is equal to 82.86%, the recall is 29 divided by 46 which is equal to 63.04%. After repeated prediction, the average accuracy of AIC model was 79.83%

Similarly, the results of confusion matrix predicted by BIC model are as follows:

Table 8. BIC confused matrix

	glm.pred	
Y.test	0	1
0	61	7
1	19	27

In this confusion matrix, the number of TP is 27, the number of FP is 7 and the number of FN is 19, precision is 27 divided by 34 which is equal to 79.41% and recall is 27 divided by 46 which is equal to 58.70%. Repeated the prediction, the average accuracy of BIC model was 77.20%.

According to the above results, AIC has higher accuracy than BIC, so AIC model is better than BIC model in this data set.

3. Conclusion

Through research, this paper finds that BIC is too sensitive to model over-fitting in the process of model selection for simple models, which leads to the loss of model prediction accuracy. AIC model is 2% more accurate than BIC model and they have similar recall. The independent variables in this paper only have linear relationships. If the interaction between independent variables is taken into account, the accuracy loss of the model will be greater if one of the key parameters is removed. The case in this paper is to predict the diagnosis of a patient, in which case the recall of the confusion matrix needs to be taken into account. Compared with AIC model, BIC model has disadvantages in terms of accuracy and recall in this experiment. This experiment suspects that AIC is a relatively more effective method for model selection in relatively simple models (with fewer parameters). However, in the case of complex models (with more parameters), BIC method is more suitable, because BIC method's more severe penalty for over-fitting will make the final model simpler. However, most of the diagnostic factors for patients have relatively fewer parameters, similar to the patient diagnostic parameters in this study. Therefore, it is not recommended to use BIC method for model selection when predicting patients' diagnosis. It is suggested that the BIC method should be studied in the future to take into account the accuracy and over-fitting when the model is small.

References

[1] George Edward Pelham Box, Norman R. Draper. (1986). “Empirical Model-Building and Response Surfaces”). John Wiley & Sons, Inc.605 Third Ave. New York, NY, United States

[2] Stoica, P.; Selen, Y. (2004). "Model-order selection: a review of information criterion rules", IEEE Signal Processing Magazine (July): 36–47, doi:10.1109/MSP.2004.1311138, S2CID 17338979

[3] Fawcett, Tom (2006). "An Introduction to ROC Analysis". Pattern Recognition Letters. 27 (8): 861–874. doi:10.1016/j.patrec.2005.10.010.

[4] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". Journal of Infrastructure Systems. 26 (1): 04019036.

[5] Trnecka, M., & Trneckova, M. (2021). Model order selection for approximate Boolean matrix factorization problem. Knowledge-Based Systems, 227, 107184.

[6] Grdenić, G., Delimar, M., & Beerten, J. (2022). AC Grid Model Order Reduction Based on Interaction Modes Identification in Converter-Based Power Systems. IEEE Transactions on Power Systems.

[7] Korobkov, A. A., Diugurova, M. K., Haueisen, J., & Haardt, M. (2021, January). Multi-dimensional model order estimation using LineAr Regression of Global Eigenvalues (LaRGE) with applications to EEG and MEG recordings. In 2020 28th European Signal Processing Conference (EUSIPCO) (pp. 1005-1009). IEEE.

[8] Tamri, A., Mitiche, L., & Adamou-Mitiche, A. B. H. (2022). A Second Order Arnoldi Method with Stopping Criterion and Reduced Order Selection for Reducing Second Order Systems. Engineering, Technology & Applied Science Research, 12(3), 8712-8717.

[9] Alfke, D., Feng, L., Lombardi, L., Antonini, G., & Benner, P. (2021). Model order reduction for delay systems by iterative interpolation. International Journal for Numerical Methods in Engineering, 122(3), 684-706.

[10] Ganguli, S., Kaur, G., & Sarkar, P. (2022). Model order diminution of MIMO systems using the delta transform method with new firefly-based hybrid algorithms. Soft Computing, 26(12), 5883-5900.