Annals of the Institute of Statistical Mathematics (2022) 74:639–641
https://doi.org/10.1007/s10463-022-00833-7

INVITED ARTICLE: THIRD AKAIKE MEMORIAL LECTURE

Discussion of Akaike Memorial Lecture 2020: Some of the challenges of statistical applications

Masayuki Henmi[1]

Received: 10 November 2021 / Accepted: 24 November 2021 / Published online: 26 May 2022

First of all, I would like to congratulate Professor John Copas for receiving the third Akaike memorial lecture award. About fifteen years ago, I worked with him as a member of his research project on meta-analysis and at that time he told me that it was desirable for research statisticians to be involved in applied works with experts of some other areas and the development of statistical methodology in a good balance. In fact, he has continued it for a long time and produced many important works, and now he is trusted by many researchers of other areas as well as professional statisticians.

This paper consists of four parts, each of which is selected from many works of Professor Copas and addresses an essential issue in statistical science. The first topic is on prediction by regression models and a role of shrinkage in the prediction. This work is motivated by his experience of real data analysis, of which the purpose is to predict the development costs of aircrafts based on their design characteristics. One of the key observations in this work is that for the outcome variable y and the fitted value $\hat{y}$,

$$E(y|\hat{y}) = K\hat{y},\ K = \frac{\hat{\beta}^T V\beta}{\hat{\beta}^T V\hat{\beta}},$$

where β, $\hat{\beta}$ are a regression vector and its estimate, respectively, and V is a covariance matrix of a covariate vector. Then, the estimate of K is given by

$$\hat{K} = \frac{n\hat{\beta}^T V\hat{\beta} - p\sigma^2}{n\hat{\beta}^T V\hat{\beta}} = 1 - \frac{1}{F}\ \left(F = \frac{n\hat{\beta}^T V\hat{\beta}}{p\sigma^2}\right),$$

where n, p and σ^2 are the sample size (of the construction sample), the dimension of the covariate vector and the residual variance, respectively. This result indicates that as the value of F (Fisher's F statistic) is smaller, the value of $\hat{K}$ is further away from one and the prediction of y with the fitted value $\hat{y}$ gets worse. In particular, unless

✉ Masayuki Henmi
henmi@ism.ac.jp

[1] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

the sample size n is not sufficiently large compared to the dimension p of the covariate vector, it is more likely to occur. This is the problem of “overfitting” in terms of prediction. In recent years, data science has become increasingly more popular and more people use regression models for prediction in various problems. However, as Professor Copas points out in the paper, many practitioners would continue to use the standard methods (like least square method) for regression models. Although penalization or shrinkage methods such as lasso and ridge regression for prediction has been gradually recognized, his message is still important now. In these days, there are more opportunities to analyze “big data” with large sample size. However, this does not mean that we do not have to care about overfitting in regression analysis. Such big data is not necessarily sampled from a single homogeneous population and it is often heterogeneous, possibly with missing data or outliers.

The second topic is on statistical inference for non-random samples. Here, the term of “non-random sample” means that the data is not necessarily a random sample from a population of interest and is possibly biased. (So, it is still assumed that there is some stochastic mechanism (probability distribution) from which the data is generated.) Although there are many kinds of problems on selectivity bias, Professor Copas focuses on selectivity bias arising from missing data, which is the best known example in statistics as he mentions. One of the motivating examples in this work is the real data, which describes the relationship between the date when the patient was entered into the trial and the duration in which the patient with the new treatment had to stay in hospital (hospitalization rate), for a clinical trial to test a new form of kidney dialysis. In order to investigate whether the decreasing trend of the hospitalization rate was caused by some selectivity bias or not, Heckman’s model was applied to this data. This model consists of two equations. One is a regression model for the outcome based on the original distribution in the case of no selection and the other is a regression model for the latent variable which controls the selection process of patients. Professor Copas considered this model under the assumption that the hospitalization rate did in fact not depend on when the patient was entered into the trial, and found that this model could explain the observed data as well as the standard regression model. As he mentions in the paper, we cannot know that which model is more correct just by looking the observed data. It is also undesirable to apply some model selection procedure because the Heckman’s model is based on an untestable assumption about the selection process of patients. His message that the assumptions for statistical analysis have to reflect the scientific context on the data is important in every case, but particularly so in the problem of non-random samples.

I would like to ask about one technical issue in the analysis with Heckman’s model. Professor Copas used a nonlinear regression model to estimate the parameters of Heckman’s model. This method is so-called the Heckman’s two-step procedure and as he points out in the paper, the two equations in Heckman’s model should have different covariates when this method is applied. On the other hand, it seems that the maximum likelihood method can be applied to estimate the parameters in this model. What happens if we apply the maximum likelihood method to Heckman’s model where the covariate of date when the patient was entered into the trial is included in both of the two equations?

The last topic is on publication bias in meta-analysis, which is another example of problems on non-random samples. Meta-analysis is typically conducted based on the data from published sources such as journals, research reports, data bases and so on. Such data are quite observational and likely to be published with some bias. This is called the problem of publication bias and is one of the most challenging problems in meta-analysis. Professor Copas's works on publication bias are also motivated by examples of real data, ranging from criminology to medical research. Among his works on this issue, 'Copas method' is the most well-known and widely used in practice. As he mentions in the paper, however, this method has some problems and he proposes a new method to overcome them. It is based on direct modeling of selection function, which is much simpler than 'Copas method'. Although this type of selection modelling approach has been studied since the early stage of research on publication bias, it seems that the t-value of 2-tail test has not be used in the literature before. The funnel plot of the criminological review seems to be well-explained by this method, but I would like to mention one issue. Professor Copas says that it is more difficult to estimate the selection parameters in the model if the systematic review has fewer studies. This means that the model for the distribution of observed data (from observed studies) is not necessarily perfectly identifiable. It is maybe because the observed data has no information on the selection process of studies, which is modeled with the parametric selection function. Even though it is possible to estimate all the parameters in the model for the example of the criminological review, I think that we should be careful because such possibility comes from the strong parametric assumption for the selection function and there are many possible choices of parametric forms for the selection function under the assumption that the selection function is an increasing function of the absolute t-value. I believe that this method is still useful, but I hope that this issue is included in "further research" , which is mentioned in the paper.

The problem of non-random samples can arise not only in missing data and publication bias, but also in other forms regarding data analysis. The research on this important problem has a long history in statistics and Professor Copas has made significant contributions on this problem. In recent years, we often have a much larger amount of data than before and machine learning methods are increasingly used to analyze it. In such a case, it would be still important to care about possible bias caused by the selection or generation process of data from a statistical point of view.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.