Annals of the Institute of Statistical Mathematics (2022) 74:649–652
https://doi.org/10.1007/s10463-022-00830-w

INVITED ARTICLE: THIRD AKAIKE MEMORIAL LECTURE

Author's rejoinder to the discussion of the Akaike Memorial Lecture 2020

John Copas[1,2]

Received: 18 March 2022 / Accepted: 24 March 2022 / Published online: 19 May 2022

Keywords Shrinkage of predictions · Selectivity bias · Model sensitivity · Publication bias

First I would like to thank the discussants, Professors Henmi and Taguri, for their kind words about my lecture, and for their general support of my theme for the need to balance our methodological research with attention to how statistical methods are being used in practice in other application areas. Of course I also agree with the comments for the need for clarity in the objectives of our analyses, and the importance of avoiding the ever-present problem of over-fitting and the loss of control of assumptions and unexplained heterogeneity.

1 Professor Henmi's comments

Professor Henmi gives a good summary of my section on Non-random Samples, and also asks a very important question—what happens if the covariate x appears in both equations in the Heckman model, both as an ordinary linear regression term in the first equation for y, and also as a linear component of the selection function in the same way as before? The problem is only mentioned very briefly in my lecture. Because of the assumptions of Gaussian errors, the likelihood function is still well-defined by the Heckman bivariate equation model. Conditioning on selection now gives a single regression of the observed y on x, with two additive components, the linear regression in the y equation as before, plus the linear regression of y on x appearing within the Mills ratio term. The difficulty is separating out the two linear regression components. The Mills ratio function $\lambda(u) = \phi(u)/\Phi(u)$ is nearly linear over a wide part of its range, and so for some data sets the two regression components may be highly collinear, and hence difficult to separate. In that case, standard

✉ John Copas
jbc@stats.warwick.ac.uk

[1] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[2] University College London, London, UK

maximum likelihood algorithms will give a fitted model with very large standard errors, showing that no useful inference is possible. Examples in the econometrics literature suggest that this a major problem, hence the advice that the use of two separate linear regression terms should be avoided.

Of course the Heckman simultaneous equations with bivariate normal residuals only works because we are assuming a probit model for selection. It might be worth exploring the possibility of using a logistic model instead, more in-line with most other applications of binary regression models. The assumptions of normality required by the Heckman approach would no longer apply, although the likelihood function is still fully defined. It would be interesting to see if similar methods would still be possible, in particular to explore the extent to which the problem of collinearity might still apply.

Again, I appreciate Prof. Henmi's comments on the final section of the lecture on publication bias in meta-analysis. He questions the strong dependence of the analysis on the particular form assumed for the dependence of publication on the value of the two-tailed test statistic $|t|$. As he says, this may not always be appropriate in practice. In fact my interest in this was suggested in a comment I received after the publication of my medical paper assessing the influence of passive smoking on lung cancer (Copas and Shi, 2000) which used my earlier model which associates publication with the usual one-tailed test statistic t. The comment pointed out that I was associating publication with large positive values of t (strong evidence that lung cancer *is* associated with passive smoking), but ignoring the fact that studies with large negative values of t (strong evidence that lung cancer is *not* associated with passive smoking) are also of interest and so are also more likely to be published. A similar comment could also be made about the criminological application in the final section of my lecture, hence the use of the two-tail test statistic $|t|$.

In statistical modeling, it is difficult (or misleading) to ask whether a model is 'correct' so that the inference based on it is also 'correct'. It is not like a laboratory experiment where we take measurements and then draw our conclusions—we can then assume our conclusions are correct if we have first checked that the measuring instrument is accurate. In statistics, I think it is more meaningful to think of an analysis as a 'sensitivity analysis' rather than a 'definitive conclusion': essentially we are saying that *IF* the model is correct *THEN* we can accept the conclusion (subject of course to the degree of uncertainty implied by the statistical method we have used). Instead of asking whether the model is 'correct', we might consider whether the model is 'reasonable' or 'sensible' in the light of the current state of knowledge of the application involved.

2 Professor Taguri's comments

As mentioned in my lecture, there have many further advances in regression methods since my early paper in 2003, and I am grateful to Prof. Taguri for mentioning some of these in his discussion. He uses the much-quoted batting averages data as a simple example of shrinkage, using the early batting averages of baseball players to predict their later season's averages. His Figure 1 shows that the James–Stein

predictor (closely related to my shrinkage factor K) substantially over-shrinks the actual values of these later averages, as remarked in the final part of the caption of this plot. As an Englishman I know nothing about American baseball so find it hard to assess what statistical assumptions might be appropriate for these data. However, my Example 1, shown in Figure 1 of my lecture, shows a similar degree of shrinkage of the validation sample using the regression approach, the vertical coordinates of the points marked as circles being much more dispersed than their predictions using the dotted line.

One of the main points about the second section of my lecture is to highlight what is perhaps the most common error in statistical applications, and that is to use a standard statistical method based on the assumption of randomization on the grounds that the data 'look random', just as the scatter plot in Figure 3 of the paper looks like a random sample from a simple linear regression model (model A). In fact model A is wrong, data looking like this could also have arisen from model B, where the association between y and x is simply the result of the way in which the sample was selected. In my earlier comment to Prof. Henmi on this example, I think we need to be cautious about describing a model as 'the truth'—in practice all we can ever hope to achieve is a sensitivity analysis, that there is a special case within Model C which predicts an association between y and x similar to that observed. Even if we can find a definitive version of Model C with this property, we will almost never be able to conclude that model C is unique. As Prof. Taguri comments, non-compliance in clinical trials is an important practical example of this discussion.

Another point to note in this example is that the Heckman analysis of model B does not just depend on the observed joint distribution of y on x, but also on the results of a separate probit analysis of the medical records of the proportions of patients who were allocated to the new treatment at the different values of x. A full analysis of the problem should consider all relevant data, including plots of the joint distribution of y and x for those patients allocated to the traditional treatment. It seems unlikely that we would be able to draw any clear inference about Model C without making full use of all available data.

The comments about the robust likelihood section are well taken—the Copas/Eguchi approach is based on standard methods in information geometry, and is just one of several attempts to develop a robust version of the likelihood function, including the semi-parametric approach mentioned by Prof. Taguri. Our paper also includes a semi-parametric version, which, in the simple example being discussed, is based on just the arithmetic mean of the logarithms of the observed incomes. In the graph of the robust log likelihoods shown in Figure 4, the semi-parametric version is shown as the dotted line.

Prof. Taguri raises two helpful points about the publication bias section of my lecture. Yes, there clearly is an identification problem with the parameter α. More recent work, starting with Copas (2013), suggests how α can be identified as a function of the marginal probability of selection, most easily interpreted in terms of the number of comparable studies which have NOT been selected for publication. This number will never be known exactly, but those working in the area in question will

often have at least some knowledge of comparable unpublished studies, suggesting at least a subjective bound for the marginal selection probability.

The final point was raised in some of the earlier papers about the 'small study effect', that skewness in funnel plots seem to show that small studies are more likely to have large treatment effects than larger studies. The size of a trial clearly has an effect on how that trial is set up and supervised, and in principal it may well be possible to add appropriate covariates into the model to allow for this. To my knowledge this has never been attempted, and Prof. Taguri is probably right in suspecting that in practice it would just lead to further technical difficulties.

Declaration

Conflict of interest Neither the author nor the discussants have any conflcts of interest to declare.

Reference

Copas, J. B. (2013). A likelihood sensitivity analysis for publication bias in meta-analysis. *Applied Statistics*, *62*, 47–66.

Copas, J. B., Shi, J. Q. (2000). Reanalysis of epidemiological evidence on lung cancer and passive smoking. *British Medical Journal, 7232*, 417–418.