Whitehead et al. Response to "Misunderstandings of Multiple Systems Estimation."

John Whitehead , James Jackson , Alex Balch & Brian Francis (2020): Whitehead et al. Response to "Misunderstandings of Multiple Systems Estimation.", Journal of Human Trafficking, DOI: 10.1080/23322705.2020.1833573

Publishers version available at
https://doi.org/10.1080/23322705.2020.1833573

We welcome this letter concerning our recent paper (referred to here as WJBF). In their second paragraph, the authors explain the value of estimating the number of victims of modern slavery in richer countries, writing that before the implementation of MSE in this context no reliable method existed. Our paper evaluated MSE, as presented by Bales, Hesketh and Silverman (2015) (BHS), and concluded that the approach is not reliable. Already the Office for National Statistics (2020) has distanced itself from MSE, as we discuss below.

The letter criticises our paper, and seeks to defend using MSE. Here we respond to those criticisms and re-emphasise our original conclusions. Using terminology defined in WJBF, our original objectives were to:

(i) explore the effects of using significance testing to select which two-way interaction terms to allow for, on the accuracy of confidence intervals for the total number of victims;

(ii) evaluate the effects of ignoring three-way interactions, as BHS do.

We examine the points made concerning our work on (i), and survey developments made since our paper. Other authors have found similar shortcomings to those we uncovered, and made partial progress towards overcoming them. The letter finds fault in our examination of issue (ii), but provides no convincing defence of ignoring these interactions. We restate our concerns and clarify how an example in WJBF supports them. Other points made by the writers are then considered and, in the conclusions section we address the potential impact on policy of our findings.

The effects of selection on confidence intervals

In discussing the treatment of two-way interactions in BHS, we join those who have applied MSE in the estimation of the number of victims of modern slavery and assume that all three way interactions are equal to 1 (in the multiplicative parameterisation of the model). Our reasons for not actually believing this are covered in the next section.

Two-stage procedures, in which the same data are used first to select a model and then to draw inferences from that model without allowance for model selection, are unsound as has been recognised in many practical settings. The simulation exercise in WJBF assessed the magnitude of this problem for MSE, and found the coverage probability of confidence intervals could be as low as 39.1% rather than the intended 95%.

It is not clear whether the authors of the letter seek to defend the use of two-stage procedures, or believe that our demonstration of this problem was flawed. They state that they "provide a high-level summary of the findings" of their investigations and promise readers "further details on the statistical aspects of the findings" upon request. They have declined to share those details with us, and so we are unable to identify which aspects of our study they are unhappy with. They claim that WJBF suggests "that model selection is to be avoided and that saturated models are to be preferred". To be clear: we suggest that model selection leads to invalid confidence intervals and that MSE is to be avoided altogether.

If three-way interactions truly were equal to 1, then fitting all two-way interactions would be a way forward, even though the wide confidence intervals that result would cast doubt on the value of the analysis. When preparing WJBF, the only mention that we found of the problem of model selection in MSE was in IWGDMF (1995). There, bootstrapping was suggested as a potential solution, although possible difficulties in applying it were highlighted. Those reservations led us to believe that the method might prove infeasible or else lead to intervals as wide as those from the all two-way interaction model.

Recently, Chan, Silverman and Vincent (2020) (CSV) have suggested alternative ways of fitting MSE models and selecting which terms to include. Like WJBF, they assess the validity of resulting confidence intervals using simulation – the only authors that we have found who do so in this context. They present just one result relating to unadjusted confidence intervals, in which the coverage rate was 61.4%. Clearly, this is well short of the target of 95% and provides corroboration of our concerns about two-stage procedures.

CSV have also developed a bootstrap approach to correct confidence intervals for model selection, although this was not mentioned in the online version of April 2019 which was available when WJBF was being prepared. For the dataset presented in BHS and examined in WJBF, Silverman (2020: Reply to discussion) quotes (to the nearest hundred) a bootstrapped confidence interval of (9,300, 14,300) – although adjustment for model selection is not mentioned in the body of that paper. This bootstrapped interval is wider that the unadjusted interval, but narrower than that from fitting all two-way interactions. However, simulations reported by CSV estimate coverage of bootstrapped confidence intervals to be 90% in a single example based on 500 replicates. Note that for normally distributed data, 95% confidence intervals are 20% wider than 90% intervals.

If the letter writers seek to defend the use of unadjusted confidence intervals following MSE with model selection, then there is now more than just our counter-example to discredit. If they seek to promote the use of bootstrapping, then we would agree that the method shows more promise than we had initially imagined. However, coverage rates closer to or exceeding 95% would need to be demonstrated in simulations with greater numbers of replicates for a wide range of "true models" before the approach could be considered reliable. Even such a demonstration – which would require substantial computing time – would fail to address the problem discussed in the next section.

The possibility of three-way interactions

Three-way interactions appear in log-linear models for counts of victims on lists. We do not know how to explain them to non-statisticians. To date, we have not seen any intuitive explanation of what three-way interactions are, and what the assumption that they are equal to 1 means. How then can investigators judge whether this assumption is reasonable? This question is especially important in Bayesian versions of the analysis. There, the prior opinion of investigators is taken to be that three-way interactions are certain to be equal to 1, which implies that – whatever the data – their posterior opinion on that matter will be unchanged. To accept such a prior opinion requires understanding of what it means. Investigators who do not accept this prior have no reason to accept the posterior conclusions of the analysis.

In Table 4 of WJBF we reported a model including all but one of the two-way interactions and two of the three-way interactions, selected using backward elimination. The resulting estimated number of victims was 5,552 with an un-bootstrapped confidence interval of (4,407, 7,485). The model has 17 parameters and a residual deviance of 2.366 whereas the BHS model has 12 parameters and a residual deviance of 16.351. The model of Table 4 is a closer fit (lower residual deviance), but is less parsimonious. Two popular criteria for trading off model fit against complexity are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), with smaller values indicating more desirable models. BHS selected models according to AIC, although Silverman (2020) makes extensive use of BIC. The model of Table 4 has AIC = 123.6 and BIC = 224.2, whereas the BHS model has AIC = 127.6 and BIC = 198.6. The former model has lower AIC and higher BIC, so justification could be given for choosing either. Rather than trying to pick a winner, we feel that the former model provides a sensitivity analysis for the latter. One estimate of the number of victims is half of the other and the nominal confidence interval of the former does not overlap with the BHS interval of (9,889, 13,063). Such a discrepancy between two reasonable applications of MSE suggests that the method itself is irredeemably flawed.

In WJBF we simulated data from a model with three-way interactions, and found that the analysis methods of BHS, based on software that they used, failed to provide acceptable estimates and confidence intervals for the total number of victims. We are pessimistic about there being a technical fix to this problem, given the limited extent of the data to be analysed. It is this impasse that leads to us continuing to believe that, despite advances in bootstrap methodology, the MSE method is fatally flawed.

Detailed responses to the points made by Vincent et al.

Their letter opens by suggesting that in WJBF we argue that "Multiple Systems Estimation (MSE) is something other than a widely accepted methodology for generating estimates about dynamic populations". We do not understand this, as in our introduction we wrote, "The BHS estimate has not only become a central reference point for the UK government's strategic thinking on modern slavery ... , it is also widely used by scholars working in the area". We are also mystified by the reference to an "industry standard".

Wide acceptance of a method is no guarantee of its validity. Science proceeds not only through successive advances, but also by periodic retreats to correct earlier errors. WJBF refers to a medical application of pre-testing for carry-over in the analysis of two-period cross-over clinical trials. For more than twenty years this was accepted procedure, and indeed it was part of an industry standard approach – the industry in question being pharmaceutical. One of us (JW) taught the method to students and set examination questions in which use of pre-testing attracted full marks. Then came the paper of Freeman

(1989) which presented calculations demonstrating that coverage of nominal 95% confidence intervals resulting from this method could fall to as little as 56%. Within the pharmaceutical industry and most medical schools and statistical courses the use of pre-testing disappeared, if not overnight, then as quickly as the word spread. The move to develop bootstrap confidence intervals within MSE analyses may be a positive reaction to our paper of a similar type, or a happy coincidence.

We are well aware of the importance of choosing appropriate, parsimonious models to represent data and draw conclusions, and two of us (JW and BF) have been doing so in various application areas for over 40 years each. Nevertheless, within the data analysed by BHS, it is hard to imagine appearance on one list being independent of appearance on others. Lack of significance of a model term does not guarantee that it does not exist or affect other estimates. It would be useful to specify what magnitude of interaction terms might invalidate estimates of population size, and to quantify the power of MSE analyses to detect them.

We do not agree that our findings would "disqualify the entire field of statistical modelling" – although we would not shirk if they did! The MSE problem is most unusual as the objective is to estimate the population size, rather than that being a given, and as the majority of the population is missing. The method is also unusual in relying upon an assumption – that three-way interactions are equal to 1 – that is so difficult to explain to non-statisticians and yet must be accepted in order to proceed. Clinical trial analysis, for example, can depend on assumptions such as the hazard of death being lowered by administration of a drug by the same proportion in young, middle-aged and elderly patients. Such assumptions can be debated between clinical investigators and statisticians in language which both parties understand. If they cannot be made, then alternatives are available, such as conducting separate trials in the three age groups. Their validity can be assessed after the study, in analyses with reasonable power.

The letter suggests that to investigate the BHS method we should have simulated from a model in which the four two-way interaction terms found to be non-significant by BHS were set to 1. That might have been an interesting study, but the BHS method should work whatever the truth. Fitting all two-way interactions led to coverage probabilities close to 95%. That is to be expected – although approximations to normality depending on sample size and parameter values might have caused inaccuracy. The finding serves as a validation that our simulations could demonstrate 95% coverage when it exists, and that lower coverage in other cases is not due to other factors.

In the Conclusions section of WJBF, we discuss non-existence of estimates. To be precise about this issue: for the simulation runs reported in Table 3, there was one case when all 2-way interactions are fitted (out of 10,000) where the MLE was non-existent, and in row (b) of Table 5, there were two. There were none in Table 3 when the BHS method was used, nor in the runs leading to rows (a) and (c) of Table 5. Excluding non-existent runs, estimated coverage probabilities are unchanged to the four decimal places reported and estimated values of κ change by just one or two victims out of eight or nine thousand. At the time of our analyses, the software could not be used to check the models fitted using backward elimination and summarised in row (d) of Table 5. There may well be many instances of non-existence amongst the 10,000 replicate data sets used in this case. Nevertheless, an analysis that gives no confidence interval is little better than one that gives a confidence interval that does not contain the truth – except of course that you know that it does not.

Conclusions

The Journal provides the writers of the letter with an opportunity to comment on this response. We would be very interested to learn how you explain the meaning of three-way interactions to non-statisticians, and how you characterise the assumption that they take the value 1. What is your experience of such discussions with policy makers? What underlies your insistence that mis-specifying these terms could not materially affect your analyses?

If the validity of this assumption could be demonstrated, then how might bootstrap methods for computing confidence intervals be further improved? What level of accuracy would you regard as sufficient? Do you accept that use of MSE without allowance for model selection is inappropriate, and that further improvement of the methodology is necessary before the truly life-affecting decisions that agencies have to make should rely on its findings?

The letter concludes with concern that our paper might have an impact on policy. This could have already happened, or it might be a coincidence that in March 2020 the UK Office of National Statistics included the following passage in Section 4 of its latest report on modern slavery in the UK.

> *In 2014, the Home Office produced an estimate of the scale of modern slavery in the UK of between 10,000 and 13,000 potential victims using a multiple systems estimation (MSE) approach. While the data and method used to estimate the number of victims were the best available at the time, we currently recommend that the method should not be repeated. This is because of several issues, including changes to the content of some data sources, the dependence on administrative data and issues surrounding the statistical model used.*

We welcome the conclusion of the ONS, and the alternative approaches suggested in their report. Getting the statistics right is essential, and either the continued use of an invalid method of estimation, or the false rejection of a valid method, would be of serious concern.

References

Bales, K., Hesketh, O., & Silverman, B. (2015). Modern slavery in the UK: How many victims? Significance, 12, 16–21. doi:10.1111/j.1740-9713.2015.00824.x

Chan, L., Silverman, B. W. and Vincent, K. (2020) Multiple systems estimation for sparse capture data: inferential challenges when there are non-overlapping lists. Journal of the American Statistical Association. doi: 10.1080/01621459.2019.1708748.

Freeman, P. (1989). The performance of the two-stage analysis of two-treatment, two-period cross-over trials. Statistics in Medicine, 8, 1421–1432.

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. American Journal of Epidemiology, 142, 1047–1058.

Office for National Statistics (2020). *Modern slavery in the UK: March 2020*. https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/modernslaveryintheuk/march2020

Silverman, B. W. (2020). Multiple-systems analysis for the quantification of modern slavery: classical and Bayesian approaches (with discussion). *Journal of the Royal Statistical Society Series A: Statistics in Society.* doi.org/10.1080/01621459.2019.1708748.