Bull. Aust. Math. Soc. **101** (2020), 342–344
doi:10.1017/S0004972719001369

THE PERFORMANCE OF SOME STATISTICAL PROCEDURES USED IN CASE-CONTROL STUDIES AND METHYLOMICS

RUPERT E. H. KUVEKE

(Received 29 October 2019; first published online 8 January 2020)

2010 *Mathematics subject classification*: primary 62F03; secondary 92D20.

Keywords and phrases: post-model-selection confidence interval, coverage probability, general regression model, DNA methylation analysis.

The use of statistical procedures is ubiquitous in scientific investigations. In many areas of applied statistics, it is common practice to conduct a data-based statistical selection procedure and then to carry out statistical inference with the selected model, using the same data. However, it is well known that using the same data for both model formulation and subsequent statistical inference can result in invalid inferences.

In the context of regression models, a common two-stage data-based selection procedure may be described as follows. Firstly, for some data set, a statistical model selection procedure, such as a preliminary hypothesis test of a parameter, is carried out to select a particular model. Then, using the same data set, a confidence interval for a parameter of interest, or a prediction interval for a random variable of interest, is constructed under the false assumption that the selected model has been provided *a priori* as the true model. This false *a priori* assumption can lead to any subsequent statistical inferences made using intervals constructed in this manner being invalid. While this invalidation is a 'well-established fact' [2, page 214], it is often ignored or overlooked in practice.

In this thesis, we assess the performance of some statistical procedures in three distinct contexts and consider the effect these procedures have on subsequent statistical inference. Firstly, we consider the use of a preliminary hypothesis test of a vector parameter, in the context of nested general regression models. Secondly, we consider the use of a preliminary hypothesis test of a scalar parameter, in the context of nested linear regression models. Thirdly, we consider the use of statistical procedures in epigenomics methods for the detection of differentially methylated genomic regions.

Thesis submitted to La Trobe University in March 2019; degree approved on 14 August 2019; principal supervisor Paul Kabaila, co-supervisor Agus Salim.

We consider both theoretical and applied contexts in this thesis due to our interest in both these fields of statistics.

This thesis consists of four chapters. In Chapter 1, we discuss some of the problems associated with making valid statistical inferences after the use of some statistical procedure. We present a review of the literature in these fields which focuses on the problems that statistical procedures may create for subsequent statistical inference. Finally, we discuss the work carried out in this thesis to further knowledge on the performances of some statistical procedures which may be used in case-control studies or in Methylomics.

In Chapter 2 we consider a preliminary hypothesis test of a vector parameter to select between two nested general regression models. To assess the effect of this procedure on subsequent inference, we derive a computationally convenient formula for the large sample coverage probability of the subsequently constructed confidence interval for a scalar parameter of interest. Previously, this large sample coverage probability could only be estimated by time-intensive simulation (see [1]). Our elegant formula requires only the evaluation of a trivial term added to at most a triple integral, regardless of the dimension of the vector parameter assessed in the preliminary hypothesis test. In addition, the computation of the large sample minimum coverage probability of this post-model-selection confidence interval using our formula only ever requires a minimisation over two scalar parameters.

Our results have two main applications. Firstly, they can be used to swiftly obtain the minimum large sample coverage probability of the post-model-selection confidence interval, which should provide a good indication of whether or not the finite sample minimum coverage probability of this confidence interval is far below its nominal level. Secondly, they can be used to swiftly identify the regions of the parameter space likely to contain the finite sample minimum coverage probability, narrowing down the regions where one would search for this coverage via simulation.

Using real case-control study data, we illustrate the practical application of our formula to a confidence interval for the odds ratio of myocardial infarction when the exposure is recent oral contraceptive use, following a preliminary test that two specified interactions in a logistic regression model are zero.

The work in this chapter was published in [3].

In Chapter 3, we consider a preliminary hypothesis test of a scalar parameter to select between two nested linear regression models. We assess the differences in the effects of this test on the coverage probabilities of a post-model-selection prediction interval for a random variable of interest and a post-model-selection confidence interval for the corresponding scalar parameter of interest. We derive expected length formulas which may be used to determine whether or not it is appropriate to use the post-model-selection prediction interval.

Chapters 2 and 3 focus on some of the theoretical aspects of post-model-selection statistical inference. In Chapter 4 we consider the performance of some statistical procedures in an applied context, namely in the identification of significant differentially methylated regions (DMRs) in genomes.

https://doi.org/10.1017/S0004972719001369 Published online by Cambridge University Press

The identification of significant DMRs is regarded as 'one of the key challenges in DNA methylation studies' [5, page 737]. Due to the importance of DNA methylation in animal and plant organisms, a large number of methods for the identification of statistically and biologically significant DMRs have been developed in recent years. These DMR identification methods are predominantly designed for use with mammalian genomes, in particular the human genome. Typically, these methods incorporate a statistical selection procedure, which may be formal or ad-hoc and utilise a general regression model to model methylation data. However, as noted by [5, page 737], 'there is no clear consensus among existing approaches' on the best way in which to identify DMRs. One clear problem, as discussed by [4, page 1], is that 'current computational approaches for detecting ... [DMRs] ... do not provide accurate statistical inference'.

We assess and compare the statistical validity of four cutting-edge epigenomics methods for the detection of genomic regions which are truly differentially methylated between two conditions, in the context of plant genomes. For our analyses we use both simulated data and publicly available genomic data from the plant *Arabidopsis thaliana*. The chosen DMR identification methods, namely BSmooth, dmrseq, DSS-single and methylSig, each incorporate a statistical procedure in their DMR identification process. We show through multiple analyses that dmrseq, created by [4], typically matches or outperforms the other methods in the identification of statistically significant DMRs, in terms of sensitivity, specificity and precision. To the best of our knowledge this is the first comparative study of these DMR identification methods in the context of plant genomes.

Acknowledgement

This work was supported by an Australian Government Research Training Program Scholarship.

References

[1] N. L. Hjort and G. Claeskens, 'Frequentist model average estimators', *J. Amer. Statist. Assoc.* **98**(464) (2003), 879–899.

[2] C. M. Hurvich and C.-L. Tsai, 'The impact of model selection on inference in linear regression', *Amer. Statist.* **44**(3) (1990), 214–217.

[3] P. Kabaila and R. E. H. Kuveke, 'The large sample coverage probability of confidence intervals in general regression models after a preliminary hypothesis test', *Scand. J. Stat.* **46**(2) (2019), 432–445.

[4] K. D. Korthauer, S. Chakraborty, Y. Benjamini and R. A. Irizarry, 'Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing', *Biostatistics* **20**(3) (2018), 367–383.

[5] A. Shafi, C. Mitrea, T. Nguyen and S. Draghici, 'A survey of the approaches for identifying differential methylation using bisulfite sequencing data', *Brief. Bioinform.* **19**(5) (2018), 737–753.

RUPERT E. H. KUVEKE, Department of Mathematics and Statistics,
La Trobe University, Bundoora 3086, Victoria, Australia
e-mail: r.kuveke@latrobe.edu.au

https://doi.org/10.1017/S0004972719001369 Published online by Cambridge University Press