SC/66b/EM/02

Progress Report and proposed work plan for further analyses of Antarctic minke whale (Balaenoptera bonaerensis) body condition data

John McKinlay, Bill de la Mare, Michael Double and Alan Welsh

INTERNATIONAL WHALING COMMISSION

Progress Report and proposed work plan for further analyses of Antarctic minke whale (*Balaenoptera bonaerensis*) body condition data

JOHN MCKINLAY [1], BILL DE LA MARE [2], MICHAEL DOUBLE[3] AND ALAN WELSH [4]

[1,2,3] Australian Antarctic Division, 203 Channel Highway, Kingston, Australia, 7050. Contact email: john.mckinlay@aad.gov.au

[4] Mathematical Sciences Institute, The Australian National University, John Dedman Building 27 Union Lane, Canberra, ACT, Australia, 2601

ABSTRACT

In this paper we reflect on why we think the question of declining body condition in minke whales remains open. We provide detail about the most important of our concerns in relation to past analyses of JARPA/JARPA II data. A simulation experiment is used to demonstrate the key point that model selection by the Bayesian Information Criterion can, in some circumstances, oversimplify a model to such an extent that it misrepresents a majority of the data on which the model is based. We conclude by setting out a plan for an analysis of JARPA/JARPA II data that would address our concerns and shed further light on this potentially important question.

KEYWORDS: ANTARCTIC MINKE WHALE, MODEL SELECTION, JARPA, JARPA II, BODY CONDITION, INFORMATION CRITERION.

INTRODUCTION

Has there been any appreciable change in Antarctic minke whale (*Balaenoptera bonaerensis*) body condition, as reflected in some of the analyses of data collected during the JARPA/JARPA II program? We are uncertain. Of that, at least, we are certain. The main source of our uncertainty stems from the fact that our analyses to date (de la Mare *et al*., 2014a,b) provide results that are equivocal on the question of declining body condition. It would seem that in some subsets of the data there may be a decline in body condition. However, in other subsets the opposite occurs; body condition improves over time. In many partitions of the data, perhaps for a majority of the data, there seems to be no change at all. Can these apparent contradictions be reconciled? We think they can. But to do so will require more work; more consideration of the ecology and biology underpinning models, more thought about the types of measures that will most accurately reflect body condition, and a better understanding about how the selection of predictors impacts our view of any signal in the data. We will outline a proposal for this work, which we hope will move the discussion beyond a narrow debate about which of two information criteria might be more appropriate for selecting models.

Previous work has highlighted several issues in relation to the JARPA/II data and the analysis methods that have so far been employed to assess body condition. These issues have included irregularities in the sampling protocol (Wotherspoon *et al*., 2014), adequacy of sample sizes (de la Mare *et al*., 2014a), statistical methods of accounting for sources of variation in the data (de la Mare *et al*., 2014b), and the appropriateness of information criterion (IC) procedures used for selecting models (McKinlay and de la Mare, 2015). These concerns have been demonstrated in concrete ways, either directly through analysis of the JARPA/II data or indirectly by simulation. Solutions to some of the issues have been implemented, in part or in whole, such as the adoption of a mixed model framework for analysis (Konishi and Walløe, 2015). However, several issues require further consideration beyond the qualitative and somewhat theoretical discussions that have been a feature of the debate to date. For example, no proper consideration has been given to the implications of the finding by de la Mare (2012) that quantities that should be time invariant have apparent trends in the JARPA data. No comprehensive quantitative work has been undertaken to demonstrate that the Bayesian Information Criterion (BIC) for model selection is delivering its assumed asymptotic behaviour. As discussed in

McKinlay and de la Mare (2015), no consideration has been given to the difficult issue of how to draw credible inferences for parameters of interest after application of a model selection procedure. There has been no attempt to establish that currently published models can adequately predict new (subsequent) data. While some of these issues may prove to be unimportant or negligible in their impact on analyses, they need to be investigated in a quantitative way through analysis of the data and by simulations structured to mimic JARPA/II conditions.

This paper details outstanding questions concerning possible trends in minke body condition. A simple simulation is presented to contrast the main issue associated with model selection using Akaike's Information Criterion (AIC) and BIC. We show that BIC, in finite sample size situations, can oversimplify a model to such an extent that actual trends are poorly estimated or concealed, while non-existent trends are attributed to subsets of the data where no structure exists. In the final section we provide a suggested blueprint for how the JARPA/II data could be analysed to help resolve some of the outstanding issues we discuss. The possibility of declining minke body condition is potentially an important, substantive question to consider. We hope the present work will stimulate further discussion about the issues involved in assessing this question, and lead to improved models that optimally reflect any signal in the data.

OUTSTANDING ISSUES

1. Model selection

a. Choice of information criterion

Different parts of the JARPA/II data contain different long-term temporal signals in respect of body condition (de la Mare *et al*., 2014b). Although perhaps not immediately apparent, discussion about how these contradictory signals should be captured in models has featured heavily in the debate thus far. This is, in essence, the argument about the choice of information criteria used for penalised likelihood methods for model selection[1] (McKinlay and de la Mare, 2015). Models selected by AIC tend to have many parameters and so are able to capture these complex features of the data. On the other hand, BIC tends to choose fewer parameters and so causes omitted effects to be "averaged over" those parameters that are retained in the model (this paper). This averaging effect has the potential to provide misleading results, and we later provide results from a simple simulation that demonstrates the problem.

b. Deficiencies of pure stepwise selection procedures

Development of statistical models is best advanced by affording as much attention to subject-matter driven theory as to empirical observation. Unfortunately, model selection by assessment of IC conducted in a purely stepwise fashion is an example of a data driven process that ignores potentially important theoretical aspects of the study at hand. Harrell (2001 pp. 56-60) convincingly reflects on the deficiencies of stepwise variable selection procedures, which include a propensity for terms in final models to have p-values that are too small, parameter estimates that may be biased high, and the potential to produce models that are biologically implausible by setting some parameters to exactly zero. For inference about individual parameters of interest, Harrell (2001) also notes that only a full-model fit will provide accurate standard errors and p-values with correct nominal coverage (although we note that this would presumably only be correct if the maximal model considered effectively captures the data generating mechanism). Berk *et al*. (2013) provide further discussion of these issues, including pointers to relevant primary literature highlighting the problems associated with post-selection inference.

In light of these deficiencies with purely step-wise approaches to model selection (including stepwise approaches based on IC), we contend that consideration of models and the choice of parameters to include in analyses should be selected based on structural hypotheses about underlying biological processes. If a biologically determined hypothesis is supported by the data then such a hypothesis should not be discarded on the grounds that simple models are "better" than more complex ones. If the complexity is well-founded in biological theory and supported by the data then it must be accommodated in subsequent analyses of the data. When theoretical information is used for this purpose, it is preferable to construct and evaluate a relatively small subset of models that are biologically meaningful, credible in light of existing information and interpretable. In our estimation, this process has not been seriously undertaken in relation to analyses of minke body condition.

---

[1] We continue to restrict our discussion of model selection procedures to penalised likelihood methods, largely due to the complexity of the JARPA/II sampling design and the need to apply statistical corrections to parameter estimates due to incomplete sampling coverage of the design space. Although not considered here, we note that cross-validation on suitably selected hold-out data remains another viable method for assessing model performance in these circumstances.

c. Parameter and Model Stability

Statistical models can, at best, offer an approximation to a system under study. How do we choose between IC variants that produce different approximations, containing different suites of covariates? How do we account for uncertainty associated with the selection of important covariates, and in the subsequent estimation of parameters of interest and statistical inferences that are conditioned on that selection process? One approach (the one we would advocate) is to assess the stability of models in response to small perturbations of the data (by bootstrapping) and to the choice of penalisation imposed on the model likelihood (of which AIC and BIC are simply two choices among many) (Müller and Welsh, 2010). The idea here is that a good model should be invariant to small changes in the data, and should be consistently selected across a range of penalisations. If different models are selected based on the assumed penalisation, we ask which structural components (covariates) do they have in common, and in what ways do they differ? This is akin to the idea of model sensitivity testing, a well-established concept within the IWC that is applied in many complex modelling situations. In fact, sensitivity testing was recommended by the JARPA II Expert Review Panel in relation to use of IC for selecting models of body condition based on JARPA/II data (IWC, 2014). To the best of our knowledge no such testing has yet been undertaken.

d. Mixed-model considerations

In the following section *Sampling Design* we set out our rationale for why mixed models (here taken to include linear and generalized linear mixed effects models, LMM/GLMM) are likely to be a necessary analysis tool when considering trends in minke body condition. In this section we briefly note that using IC-like approaches for mixed model selection is not straightforward and remains an area of active research (see Müller *et al*., 2013 for a recent review). As summarised by Jiang *et al*. (2008), major concerns regarding use of IC for mixed models are how to determine effective sample sizes in the presence of random effects, and how to overcome the need for likelihood functions when these may not be available (e.g. when normality cannot reasonably be assumed). Inference about individual terms in (G)LMM is similarly problematic since determining effective degrees of freedom for reference distributions is not clear-cut when random effects are present (Kenward & Roger, 1997). Parametric bootstrapping under full and reduced models can provide informative tests of individual parameters, a point we consider further in the section *A plan of Work*. We conclude by noting that little regard for mixed model selection or inference has thus far been demonstrated in any analyses of JARPA/II data (including our own).

2. Sampling Design

The JARPA/II sampling design has changed appreciably through time, both in respect of its spatial coverage within and between seasons, and in the distribution of within-season temporal coverage of sampling (see Figure 3, Wotherspoon *et al*., 2014). It has been demonstrated by simulation that it is possible that some apparent trends in body condition have been introduced as a result of non-balanced sampling in respect of space and time (Wotherspoon *et al.,* 2014). This is a crucially important point: to be able to detect trends in body condition a model must be able to correctly separate out, from any long-term trend, the within-season trend in accumulation of body condition that occurs each feeding season. How might this be achieved when the spatio-temporal coverage of the sampling design has changed over time?

One potential approach is to analyse data in a mixed model framework, accounting for these changes to the sampling regime by partitioning elements of variance between several random effects. This requires that some variables utilised in the analysis *remain* in the analysis in order to achieve statistical corrections to the actual parameters of interest, in this case the long-term trend in body condition. Unfortunately, unmindful application of model selection procedures, regardless of the IC used to facilitate the process, takes no account of these structural variables. They have the potential to be dropped from a model regardless of the necessary role they have in adjusting for unbalanced and incomplete sampling. JARPA/II data were collected according to an area stratified design based on transects, and these structural design variables should be protected during model selection to ensure they are not improperly removed. Sampling effort varied by year and strata, and many variables are subject to random effects. Variables about which specific structural/biological hypotheses are a-priori expected, and confirmed by the data to be statistically significant, should also be protected.

3. What is an appropriate measure of condition?

Analyses to date have concentrated on blubber thickness at particular body locations as the response reflecting animal condition, although fat weight has also been suggested as a potentially informative measure. We have reservations about the suitability of either of these variables for reflecting overall animal condition. The potential issue with blubber thickness as an indicator of condition is that there is considerable variability in the location and degree of deposition of blubber around the body, some of which is likely independent of condition. In a correlative study assessing potential minke whale body condition indicators, Konishi (2006) demonstrated that correlations between blubber deposition and

time in the feeding season ranged from 0.01 to 0.6, depending on the body location considered. These results informed subsequent analyses of blubber thickness for JARPA/II data, which concentrated on blubber thickness at body locations showing the highest correlation with seasonality. Our concern is that a correlation of 0.6 implies that there is still considerable variation in blubber thickness that is unaccounted by seasonality and therefore potentially unrelated to body condition. This extra variation might be explained by considering a model-based approach to determining body condition, a point we return to shortly. In passing, we note that several correlations presented in Konishi (2006) utilise indices that have common components, a practice that is recognised as having the potential to induce spurious correlation (Kronmal, 1993).

Our concern with fat weight as an indicator of condition is that the measurement is only available for a relatively small subset of the total number of animals sampled (< 15%). Given the sparse temporal and spatial coverage of *all* the available data, we would be reluctant to base any substantive conclusions upon analyses of only 15% of this. Additionally, the fat weights reported are not sufficient to provide the energy required for minke whale migration and reproduction (de la Mare, this meeting), i.e. there must be other important stores of energy, and their relationship to the measured fat weight is currently unknown.

In light of these limitations, we suggest also examining a third, complementary measure of condition, namely total body weight (conditioned on an estimated power of length). Total body weight, available for approximately 80% of the animals sampled under JARPA/II, has the potential to carry a strong signal of body condition without the assumptions inherent in considering blubber thickness, or the small samples sizes associated with blubber weight. Simply put, if two animals have the same length but different weights, the heavier specimen is considered to be in better condition that the lighter[2]. Methods for the analysis of length-weight relationships are well-established in fisheries science (Froese, 2006), and may provide a more useful alternative to simple correlative approaches.

As a final note on this point, *all* indicators of body condition should theoretically extract the same signal from the data, and concordance between different measures should provide reassurance that any trend has not simply arisen by chance. If several indicators provide inconsistent results then these should be investigated to determine the source of the inconsistency, a process that is likely to shed more light on an analysis problem than would investigation of just a single measure.

4. Access to data

In 2015 the Scientific Committee recommended that:

> *Given earlier recommendations by the Committee and the continuing debate of how best to model the data, the Committee recommends that additional analyses be undertaken on both the blubber thickness and body fat data. It encourages the various scientists involved in these analyses to collaborate to develop a set of models that best capture the Committee's previous recommendations, taking into account the structure of the underlying processes giving rise to the data. To facilitate this, the Committee suggests that the interested scientists apply for access to the data under Procedure B of the Data Availability Agreement. It requests the data holders to consider such requests favourably.*
>
> Section 13.3.1, IWC/66/Rep01 (IWC, 2015)

Accordingly, on 30 June 2015 McKinlay and de la Mare applied to the IWC Data Access Group (DAG) for access to the JARPA/II data. These are the data to which we have previously been provided access, and on which all our previous analyses are based. After 8 months of negotiation, including a face-to-face meeting in Tokyo, we have been offered access to data from only the JARPA period, excluding JARPA II. No clear rationale for withholding data from the JARPA II program has been provided. We note, and have argued to the ICR, that analyses of (only) JARPA data would provide results that are *incomparable with our previous analyses* of these data, including all results presented so far to the SC. To refuse access to data after 2005 seems a restriction that is not scientifically justified.

A SIMULATION EXPERIMENT

In this section we present results from a simple simulation that demonstrates the issue associated with 1a) above, namely how AIC and BIC deal with mild to moderate interaction effects. The generating model is an abstracted simplification of the JARPA/II design, in which a linear regression relationship is cross-classified by two categorical factors. Random

[2] This discounts differences in weight loss due to loss of blood or body material due to harpooning. Losses of this kind could be modelled if such data were available.

effects are not considered, and, unlike the JARPA/II design, data are generated to be fully balanced. Following Wilkinson and Rogers (1973) convention for expressing model formulae, we generate a design that satisfies:

$$Y \sim X * A * B$$

where Y is the response, X is a numeric covariate representing year (taking values 1-24), and A and B are factors each with 6 levels. An intercept term is implicitly present. Here, the asterisk (*) denotes crossing between terms, so the model contains main effects in X, A and B, as well as their interactions up to third order. We hereafter refer to cells in this A-B cross-classification of factors simply as 'cells'. At each design point, five replicates are generated with Gaussian error having mean 0 and sd=0.2. We induce a linear relationship between y and x in certain cells of the design, such that:

$$y_{ijk} = 4 + r_k x_{ki} + e_{ijk} \qquad i = 1, \ldots, 24; \quad j = 1, \ldots, 5; \quad k = 1, \ldots, 36.$$

where there are i design points in x, j replicates at each design point, and k cells. The rate parameter, r, is set to 0.015 for (on average) 10% of the 36 cells, and 0 for the remaining 90% of cells. The sign of r is randomly chosen so that, on average, half are positive and half are negative. In other words, around 5% of the 36 cells have a positive slope, 5% have a negative slope, and the remaining 90% of cells have randomly generated data around an overall mean of 4 (i.e. the specified intercept). While clearly an abstraction, the simulated slope, error, sample sizes and number of factor levels considered is consistent with our past experience of models fitted to JARPA/II data.

A fully parameterised model of the simulated data was then subjected to step-wise selection using stepAIC from the MASS package (Venables and Ripley, 2002), applying an appropriate penalisation for each of AIC and BIC. In expanded notation, using ':' to indicate interactions, we fit a model with terms for X, A, B, X:A, X:B, A:B, and X:A:B. It is worth noting that this can be considered the maximally correct model, in the sense that it must always contain the true data generating model, even in the event that a simpler model is more appropriate for any individual realisation of the simulation. Potential model misspecification is therefore not a factor in what follows. The maximal model is then subjected to stepwise selection (using both forward and backward steps, as necessary) with a restriction that the marginality of lower order terms is respected. Predictions and approximate 95% confidence intervals were calculated for four representative final models, and these are presented graphically along with the simulated data values used to fit the models. All code was developed in the R language for statistical computing, version 3.2.4 (R Core Team, 2016) and utilised packages MASS (Venables and Ripley, 2002), lattice (Sarkar, 2008) and latticeExtra (Sarkar and Andrews, 2013).

While undertaking this work we examined models fitted to a large number of simulated data sets (> 100). However, we have made no attempt to describe average behaviour in this simulation experiment since to do so would involve considerable complexity; one would have to account for the position of the significant slopes in the cells of the cross-classification of factors, as well as discern how simulation variables (i.e. slope size, regression error, number of factor levels, replication, number of cells with significant effect) interact with cell position. While this work would certainly be of interest, our preferred approach is to show some representative examples of the final models selected by AIC and BIC, and reflect on how these models compare with the simulated data. The model fits we present here are typical of the wider set of results obtained, a claim easily confirmed by using our data generation code (Appendix A).

As a footnote, we additionally assess the degree to which our chosen representative models change by considering 18 instead of 24 time points. This aspect speaks directly to the suggestion that further analysis should be restricted to data collected under JARPA and exclude data collected under JARPA II.

How well do AIC & BIC recover interactions?

Results from our simulations showed that AIC performed well. It correctly recovered known model structure in all cases considered, including the four examples presented in this section (Figures 1a, 2a, 3a and 4a). The downside of this success was that final models were close to, if not always, saturated and so contained a large number of parameters. We noted in previous work that AIC will tend to retain terms with a notional significance level of around 0.15 when removing single variables in nested model situations (McKinlay and de la Mare, 2015). To counter any arguments about this level being too permissive by conventional frequentist standards, we repeated the model selection process using stepAIC with a modified penalty multiplier of 3.84 (instead of 2, assumed by AIC), determined as the value defining the 95th percentile of a χ^2_1. This provides a more conservative test for exit and entry of terms, approximating a significance level of 0.05 (Venables and Ripley, 2002). This occasioned some slight simplification to a small subset of the models considered (results not shown), but, for the most part, these models were largely unchanged in comparison to their AIC counterparts, while remaining appreciably different from final models selected by BIC. By way of comparison, BIC

assumes a penalty multiplier on the number of free parameters of more than 8 for the simulated sample sizes shown here (i.e. log(n) = log(6*6*24*5) = log(4320) ≈ 8).

Results for BIC final models were mixed, but broadly fell into four model types (I-IV). In Type I models a small number of simulated trends could be effectively ignored in the final model selected. These models estimated no trend across the entire design. Figure 1b shows an example of this behaviour, where three cells in the A x B cross-classification had simulated negative trends but were ignored in the final model. Estimated regression parameters for these cells do indicate slightly different intercepts (the omnibus test of main effect of A is significant), though these are hardly noticeable.

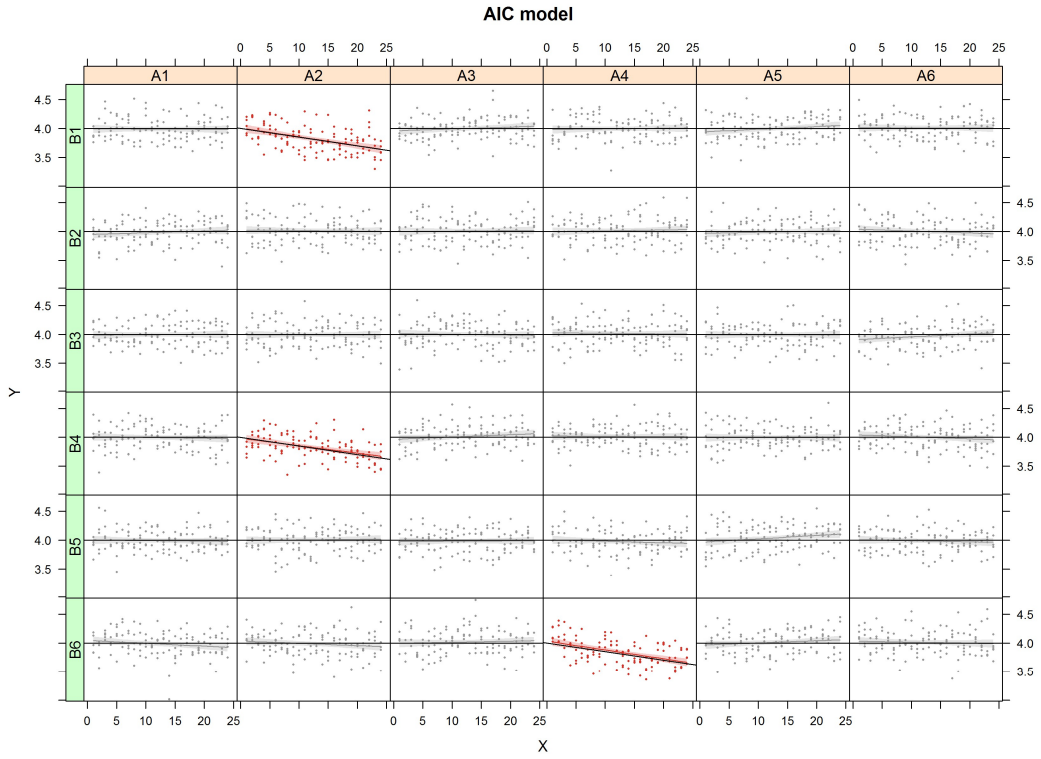
Type II model selection behaviour exhibited by BIC showed a propensity for final models to incorrectly attribute changes in slope to changes in intercept (Figure 2b). In all examples of this kind considered, the sign of the change in intercept correctly reflected the sign of the simulated slope, however the slope effect itself was effectively ignored.

An example of a Type III final model selected by BIC is shown in Figure 3b, in which a relatively small number of cells (here, 5/36) with a consistent trend are modelled as a single average trend across the entire design. Two serious deficiencies are immediately apparent: a spurious negative trend is imposed on 86% (31/36) of the data, and the estimated slopes for the five cells that have simulated negative slopes are badly biased away from their true value (i.e. the regressions for these cells clearly under-fit the data). The estimated negative slope is slight (-0.0025) but highly significant ($p<0.000001$). It is interesting to note that while 3 cells that had negative slopes were effectively ignored under BIC selection (Figure 1b), increasing the 'signal' to just 5 cells results in an entirely different model and an interpretation that is inconsistent with the majority of the data (Figure 3b).

Figure 4, our final example, shows a Type IV final model selected by BIC in which the interaction between X and a single categorical variable is retained in the model (Figure 4b). This model requires a more nuanced interpretation, but its impact on our understanding of the data is no less pernicious compared with previous examples. In this case the interaction between X and A is retained and has the effect of fixing average slope values (across all levels of B) for each level of A. Consider the first column of Figure 4b, corresponding to A1. We see that the fitted regression line for cells B1-B6 in column A1 have a slope value that is determining as an average of the slope values for all six cells in the column; that is, it is an average of two real slopes (B3 and B4) and four zero slopes (B1, B2, B5, B6). A non-existent negative relationship between Y and X is estimated for [A1; B1, B2, B5, B6], and the known negative relationship in [A1; B3, B4] is underestimated. Similar patterns can be found across other levels of A, in which one or two cells with signal are averaged across all the six cells representing levels of B.

Can these final models selected by BIC be considered reasonable representations of the underlying data? When non-existent relationships are assumed for large proportions of the data, and when actual (simulated) relationships are poorly estimated, we think the answer is self-evidently 'no'. These poor results for BIC models arise due to two effects impacting calculation of the penalty term: reasonably large sample sizes (in this case, ~4300), and the large number of degrees of freedom associated with interactions between factors. In practice, these combined effects serve to exclude minor to moderate interactions under BIC. In many cases, this results in an averaging of localized effects in higher order terms across lower-order terms in the model, inviting us to draw generalisations that don't always seem sensible in light of known, simulated structure.

a.



b.

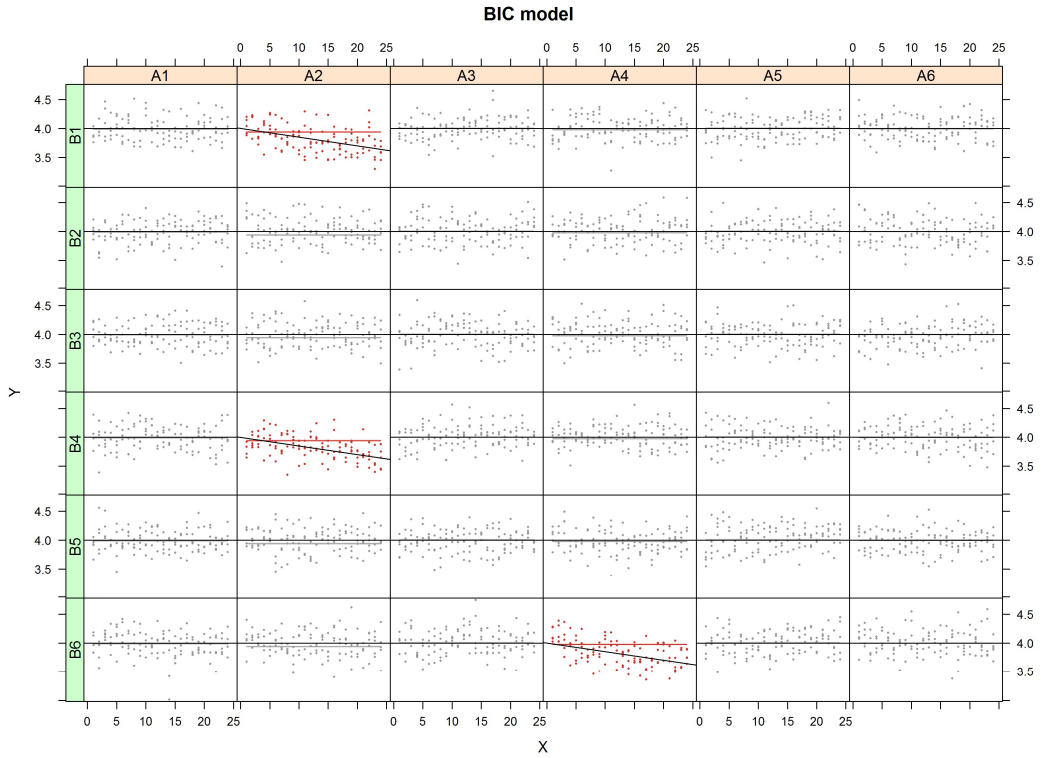
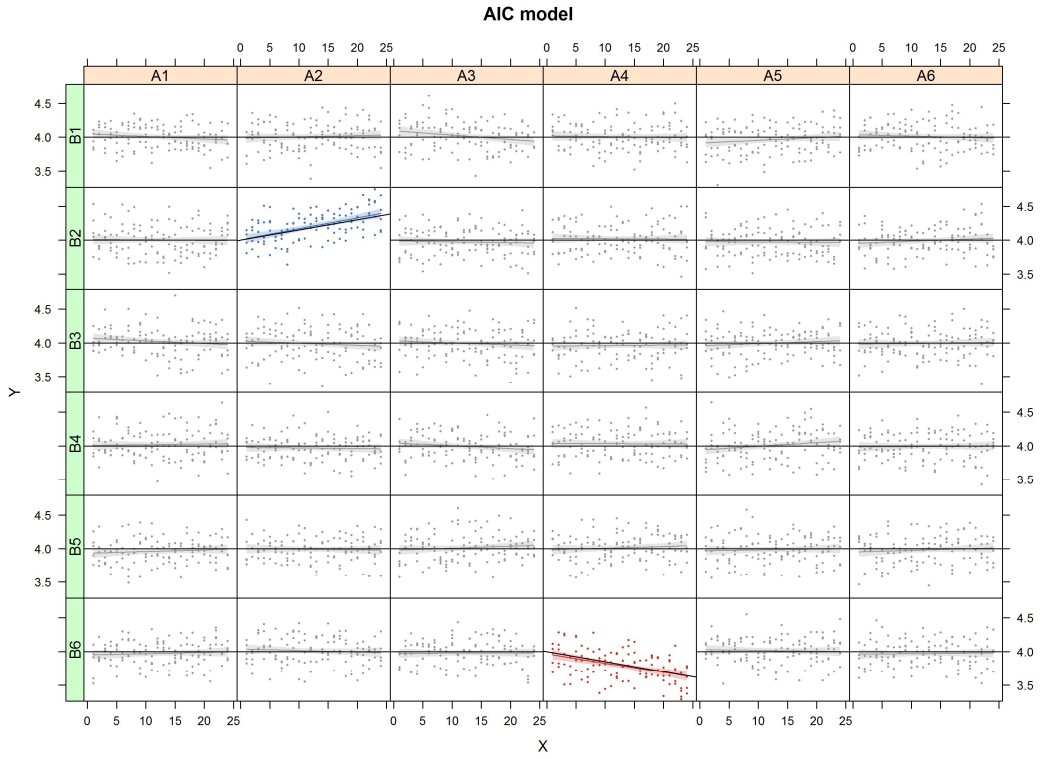


Figure 1. Model Type I: no trend effect. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.

a.



b.

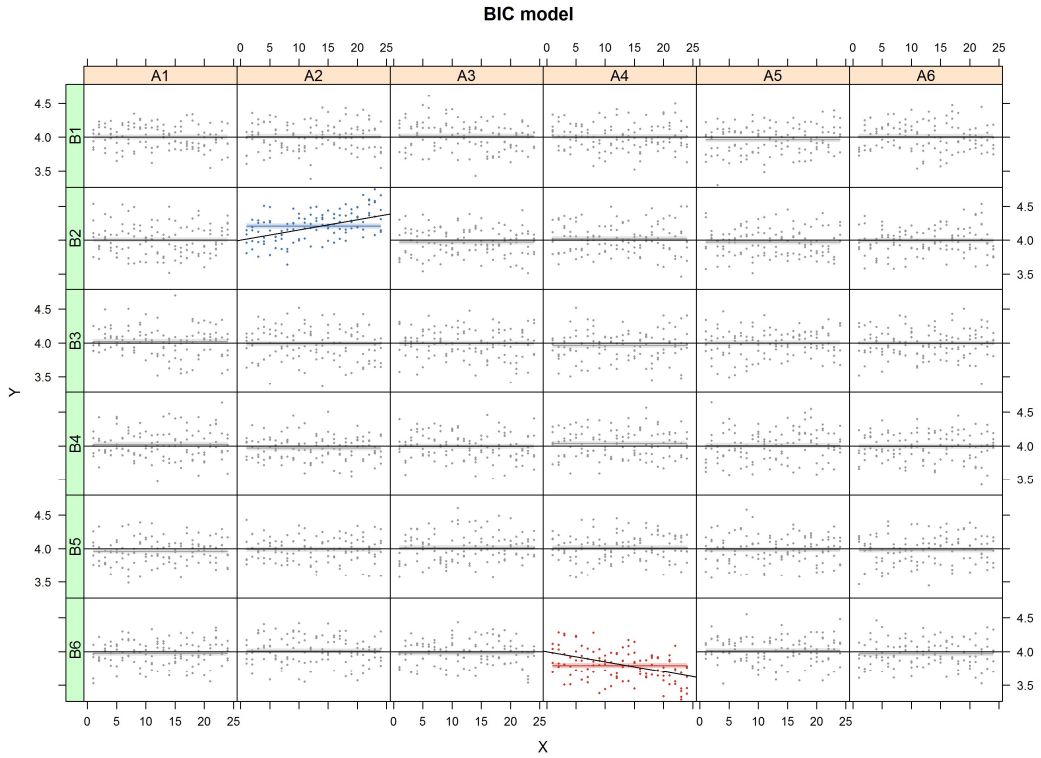
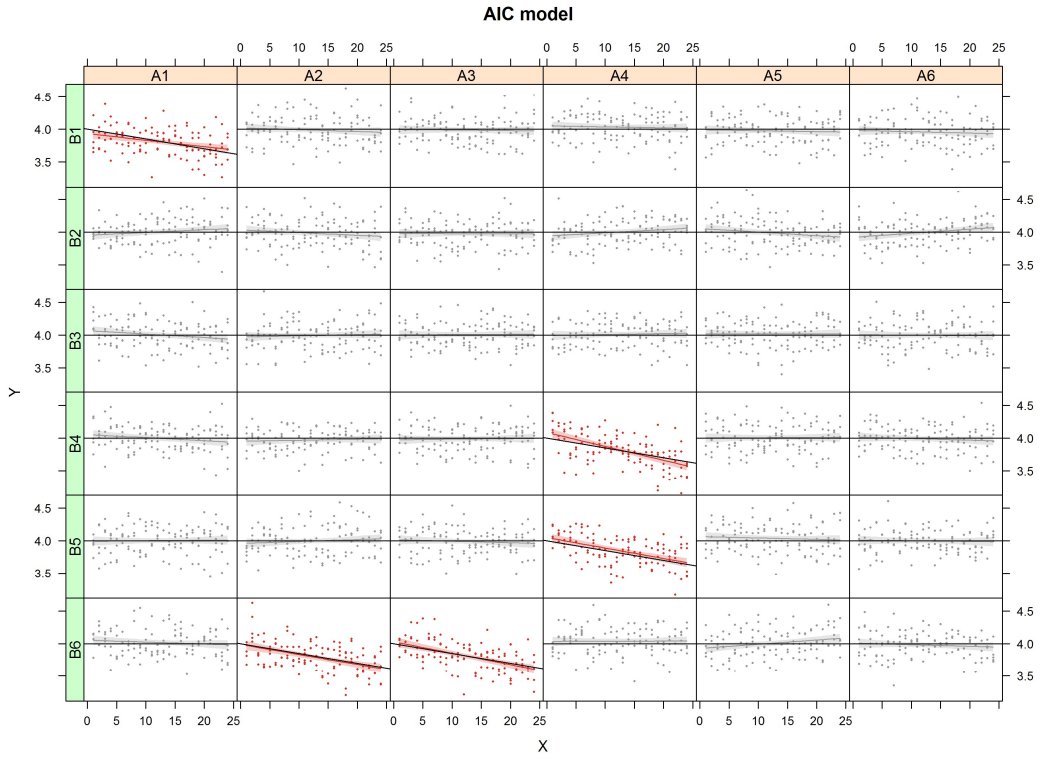


Figure 2. Model Type II: shift of intercept. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.

a.



b.

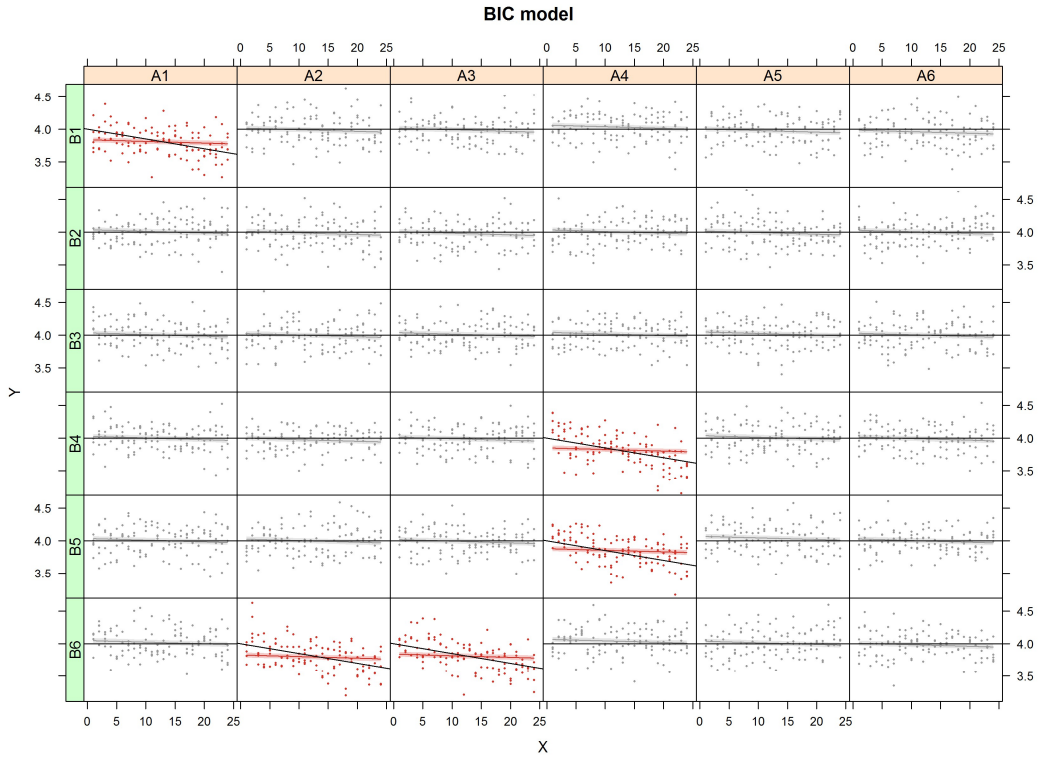
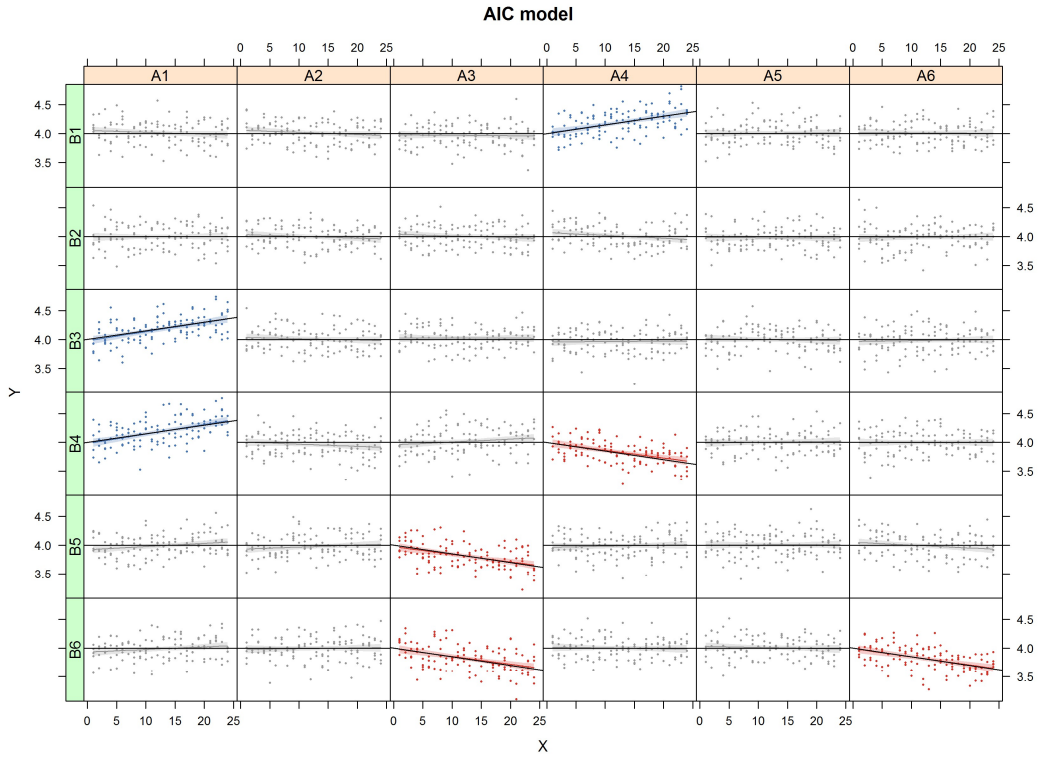


Figure 3. Model Type III: fixing slope. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.

a.



b.

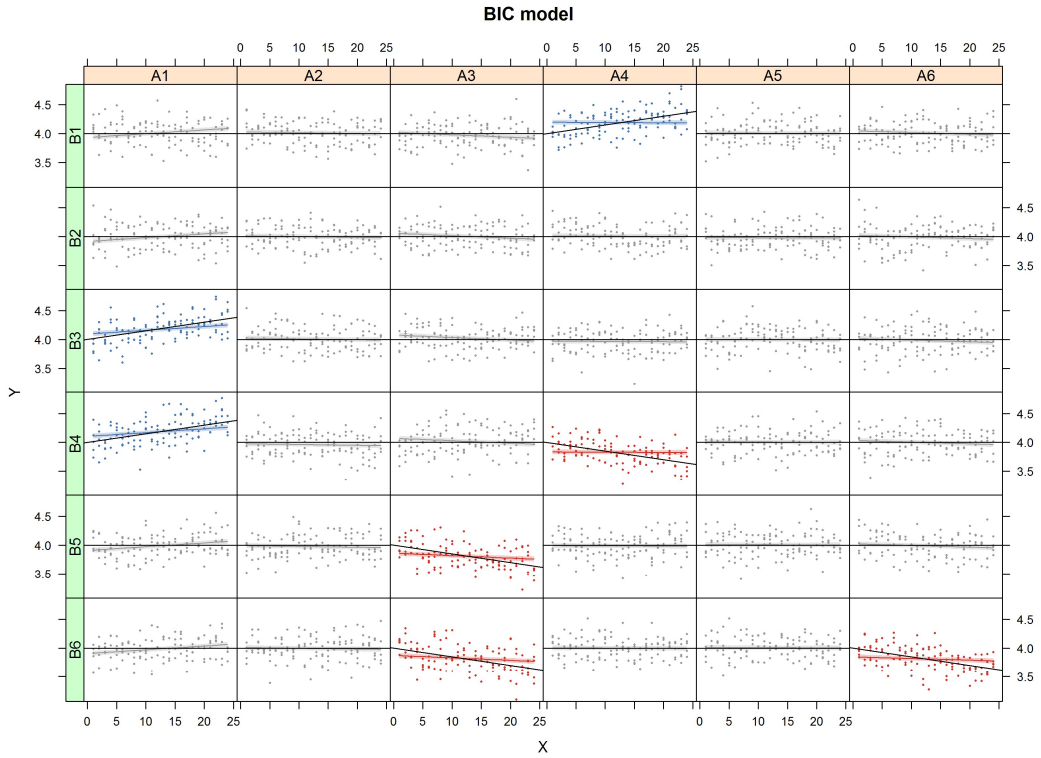


Figure 4. Model Type IV: row and column constraints. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.

Does considering 18 instead of 24 time points make a difference?

We additionally conducted model selection by AIC and BIC on a subset of the four data sets presented in the previous section by taking X in the range 1-18 rather than 1-24. Except for this truncation to consider a shorter time-series, all values for the response and covariates were exactly the same (i.e. data were not re-simulated). We undertook this exercise to assess the degree to which reduced-data models might change compared with their full-data counterparts, a point that speaks directly to the idea that further analyses should be restricted to JARPA rather than JARPA/II data.

We show that modelling 75% (18/24 years) of the data can, in some circumstances, lead to appreciably different models compared with an analysis of the full dataset. In Appendix B we present figures showing the predicted structure for the best models selected by AIC and BIC for the reduced data (Figures B1-B4). As the underlying simulation model was identical for both the full and reduced data, we expected that reduced-data models would be very similar to their full-data counterparts. To our surprise, in most cases final models were different from those selected from fits to the full data. Some of these differences were slight (e.g. differences in intercept evident in Figure 1b are dropped in the reduced-data model), but for some they were appreciable (e.g. fixed slopes within levels of factor A in Figure 4b change to fixed changes in intercept in the reduced-data model). We do not wish to belabour these results overmuch, except to say that, for these data, reducing the available dataset by 25% did make a difference to final models for both AIC and BIC. On the basis of this work, we would question the usefulness of any further analyses that utilise only JARPA data.

A PLAN OF WORK

We have tried to highlight what we feel are a number of unresolved issues in relation to the use of IC-like approaches for determining sensible models of minke body condition. Here we propose methods and approaches to analyses designed to address and overcome these issues.

1. The choice of IC used for model selection, sampling design considerations, and parameter and model stability can be considered concurrently through the use of computer intensive, resampling methods. This approach would consider parametric, semiparametric and nonparametric bootstrapping methods for GLMMs (Halekoh and Højsgaard, 2014; Davison & Hinkley, 1997; Shang & Cavanaugh, 2008). It is intended that these methods are used for both model selection and assessing the stability of selected models (sensu Müller and Welsh, 2010), as well as to develop robust tests for individual parameters of interest. The methods of Koller (2013) may also be considered. One challenging aspect of this work will be to develop a robust resampling procedure that takes into account the variation in the sampling design and the auto-correlated nature of the data. It is anticipated that some of these methods will be computationally intensive due to relatively large sample sizes, and that results may need to be presented in a staged approach over several IWC meetings.
2. Examine an extended range of mixed model diagnostics to explore whether fitted models appropriately describe the underlying data, and to determine which data (if any) are being modelled poorly. Examples of severally potentially useful diagnostics are presented in Loy and Hofmann (2014), but others may be developed as necessary (e.g. influence diagnostics of design-based model components). Lack of fit diagnostics may suggest non-linear effects or breakpoints, and these will be investigated as required. Part of this work will involve an investigation of marginal effects to isolate partitions of the data that may carry different signals, and to determine (if possible) what might be causing these differences. A range of approaches to fitting models may be considered.
3. In addition to blubber thickness and blubber weight, we propose to consider other possible indicators of body condition such as total weight (corrected for length). Where signals between models based on different condition indicators differ, explore how those differences manifest and whether they can be explained using existing data.

CONCLUSION

In this paper we have set out the reasons why we remain uncertain about the veracity of reported declines in minke body condition over the JARPA/II period. We believe further analysis is required to unambiguously assess the extent of any decline, and we have outlined a plan of work that we believe would achieve that goal. This is work we are willing to undertaken, however we require access to the data in order to do so. It is unfortunate that we have so far been unable to gain access to the JARPA/II data, in spite of our best efforts.

JARPA/II data are collected, analysed and published under the auspices of scientific whaling, purportedly for the advancement of the scientific management of whale stocks. If results from currently published work on minke body condition are to be used to inform scientific questions of interest within the IWC, then these results should be subject to

scrutiny and possible falsification. This can only be practically achieved if JARPA/II data are made available so that the reliability of results can be tested and verified. We hope the present work represents a first step toward that outcome.

REFERENCES

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. 2013. Valid post-selection inference. *The Annals of Statistics* 41(2): 802–837.

Burnham, K.P. and Anderson, D.R. 2002. *Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed*. New York: Springer. 488p.

Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap Methods and Their Application*. Cambridge University Press. 581p.

de la Mare, W.K., Candy, S., McKinlay, J., Wotherspoon, S. and Double, M. 2014a. JARPA II sample size calculations ignore spatial and temporal variability. Paper presented to the IWC review of JARPAII. SC/F14/O07.

de la Mare, W.K., Candy, S., McKinlay, J., Wotherspoon, S. and Double, M. 2014b. What can be concluded from the statistical analyses of JARPA/JARPA II body condition data? Paper presented to the IWC review of JARPAII. SC/F14/O06.

de la Mare, W.K. 2012. Lurking variables and the interpretation of statistical analyses of data collected under JARPA. Paper SC/63/O16 presented to the IWC Scientific Committee, June 2012, Panama City, Panama. 65pp.

Froese, R. 2006. Cube law, condition factor and weight-length relationships: history, meta-analysis and recommendations. *Journal of Applied Ichthyology* 22: 241–253.

Halekoh, U. and Højsgaard, S. 2014. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. *Journal of Statistical Software* 59: 1–32.

Harrell, F.E. 2001. *Regression Modeling Strategies - with Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer. 568p.

IWC 2014. SC/65b/Rep02: Report of the expert workshop to review the Japanese JARPA II Special Permit research program, Tokyo, 24-28 Feb 2014.

IWC 2015. Report of the Scientific Committee. San Diego, CA, USA, 22 May – 3 June 2015. IWC/66/Rep01.

Jiang, J., Rao, J.S., Gu, Z. and Nguyen, T. 2008. Fence methods for mixed model selection. *The Annals of Statistics* 36: 1669–1692.

Kenward, M.G. and Roger, J.H. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53: 983–997.

Koller M. 2013. Robust Estimation of Linear Mixed Models. PhD Dissertation No. 20997, University of Zurich, Switzerland. 123p.

Konishi, K. 2006. Characteristics of blubber distribution and body condition indicators for Antarctic minke whales (*Balaenoptera bonaerensis*). *Mammal Study* 31:15-22.

Konishi, K. and Walløe, L. 2014. Detailed responses to some of the conclusions and recommendations from the Review Panel regarding body condition trend in Antarctic minke whale. Paper SC/65b/EM02 presented to the IWC Scientific Committee, May 2014, Bled, Slovenia.

Kronmal, R. A. 1993. Spurious Correlation and the Fallacy of the Ratio Standard Revisited. *Journal of the Royal Statistical Society. Series A* 156(3): 379–392.

Loy, A. and Hofmann, H. 2014. HLMdiag: A suite of diagnostics for hierarchical linear models in R. *Journal of Statistical Software* 56: 1-28.

McKinlay, J.P. and de la Mare, W.K. 2015. Why BIC is not always (and may never be) an appropriate criterion for selecting terms in complex ecological models. Paper SC/66a/EM/01 presented to the IWC Scientific Committee, May 2015, San Diego, USA.

Müller, S., Scealy, J.L. and Welsh, A.H. 2013. Model Selection in Linear Mixed Models. *Statistical Science* 28: 135–167.

Müller, S. and Welsh, A.H. 2010. On Model Selection Curves. *International Statistical Review* 78: 240–256.

R Core Team 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sarkar, D. 2008. *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5. 265p.

Sarkar, D. and Andrews, F. 2013. latticeExtra: Extra Graphical Utilities Based on Lattice. R package version 0.6-26. https://CRAN.R-project.org/package=latticeExtra

Shang, J. and Cavanaugh, J.E. 2008. An assumption for the development of bootstrap variants of the Akaike information criterion in mixed models. *Statistics & Probability Letters* 78: 1422–1429.

Venables, W.N. and Ripley, B.D. 2002. *Modern applied statistics with S,* 4th Ed. Springer. 495p.

Wilkinson, G.N. and Rogers, C.E. 1973. Symbolic description of factorial models for analysis of variance. *Applied Statistics* 22:392-399.
Wotherspoon, S., Double, M.C., McKinlay, J., Candy, S., Andrews-Goff, V. and de la Mare, W.K. 2014. JARPA and JARPA II cannot monitor trends in the Antarctic ecosystem due to flawed sampling strategies. Paper presented to the IWC review of JARPAII. SC/F14/O05.

APPENDIX A – SOURCE CODE

The R code below generates the data used for the four representative models presented in the paper.

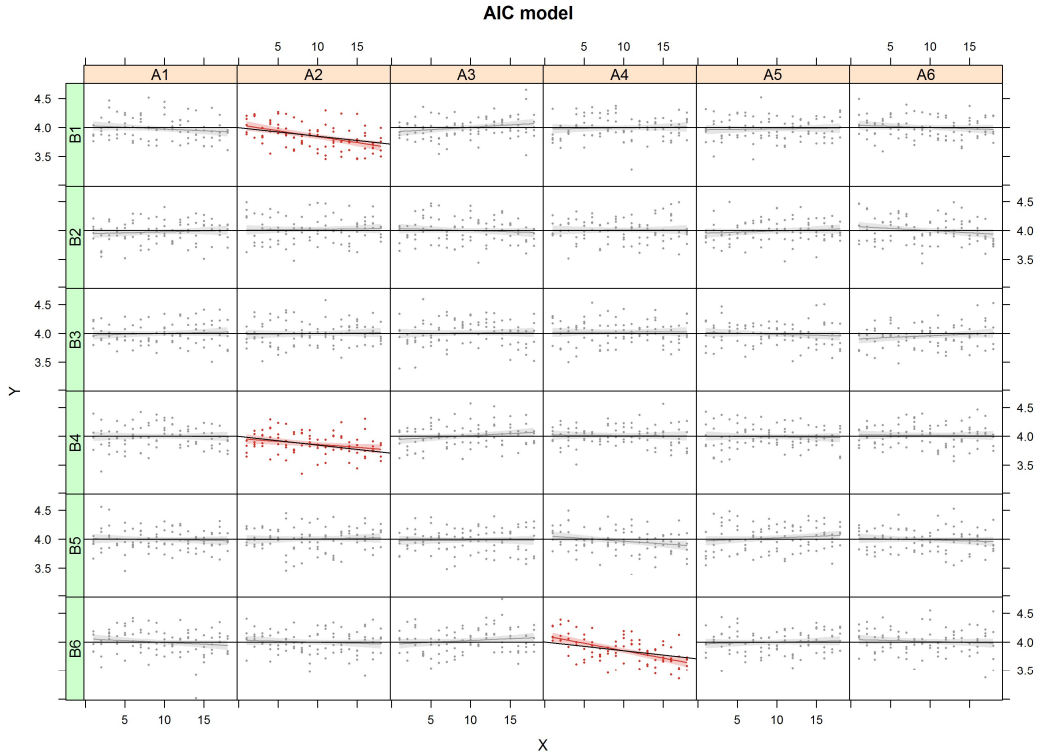
```
makedata <- function(nobs=24, X=1:nobs, nreps=5, strata=list(A=1:6, B=1:6), trend=0.015, ptrend=0.1,
direction=c("either", "up", "down"), noise=0.2, subX=NULL, seed=NULL) {
#### Description:
## Creates data for regression y~x cross-classified by (up to) two factors A and B, each with
## an arbitrary number of levels. Replicates at each design point are supported.
#### Args:
## nobs        -      integer; nbr of obs in time-series
## X           -      integer; annual time-series
## nreps       -      integer; number of replicates at each time point 1:nobs
## strata      -      list; named list of stratification variables
## trend       -      numeric; annual rate of increase
## ptrend      -      numeric; probability of trend occurring in any cell
## direction-  char; what direction for trend?
## noise       -      numeric; noise, sd for rnorm()
## subX        -      integer; subset simulated data to this range in X (e.g. subX=1:18)
## seed        -      integer; set random seed to this
#### Return value:
## dataframe of simulated data with attributes 'seed' and 'call', columns of which
## are X, reps, A, B, FA, FB, slp, Y. Note that FA & FB are factor representations of
## A & B, and slp is a binary indicator denoting a non-zero slope has been generated
## in that cell of the design
####
  if(!length(strata) %in% 1:2 && class(strata) != "list") stop("strata must be a list of length 1 or
2.")
  direction <- match.arg(direction)
  if(is.null(seed)) seed <- sample(1:10^6, 1)
  set.seed(seed, kind = NULL, normal.kind = NULL)
  vars <- c(list(X=X), list(reps=1:nreps), strata)
  d <- expand.grid(vars)
  strataF <- as.data.frame(t(apply(d[,names(strata)], 1, function(x) paste0(names(strata), x))))
  names(strataF) <- paste0(names(strata), "F")
  d <- cbind(d, strataF)
  d$Y <- d$slp <- NA
  d <- d[with(d, order(A, B, X, reps)),]
  nstrata <- length(strata)
  for(a in strata$A) {
    if(nstrata > 1) {
      for(b in strata$B) {
        beta1 <- ifelse(runif(1) <= ptrend, trend, 0)
        beta1 <- beta1 * switch(direction, up = 1, down = -1, sample(c(-1, 1), 1))
        d$slp[d$A==a & d$B==b] <- beta1
        d$Y[d$A==a & d$B==b] <- 4 + d$X[d$A==a & d$B==b]*beta1 + rnorm(nobs*nreps, 0, noise)
      }
    } else {
      beta1 <- ifelse(runif(1) <= ptrend, trend, 0)
      beta1 <- beta1 * switch(direction, up = 1, down = -1, sample(c(-1, 1), 1))
      d$slp[d$A==a] <- beta1
      d$Y[d$A==a] <- 4 + d$X[d$A==a]*beta1 + rnorm(nobs*nreps, 0, noise)
    }
  }
  if(!is.null(subX)) {
    d <- d[d$X %in% subX,]
  }
  attr(d, "seed") <- seed
  attr(d, "call") <- match.call()
  d
}

fig1 <- makedata(seed=6)
fig2 <- makedata(seed=3)
fig3 <- makedata(seed=26)
fig4 <- makedata(seed=19)
```

Subsetted data used for figures in Appendix B are obtained by adding the argument subX=1:18 to the function calls above.

APPENDIX B – RESULTS FROM 18 INSTEAD OF 24 TIME POINTS

a.



b.

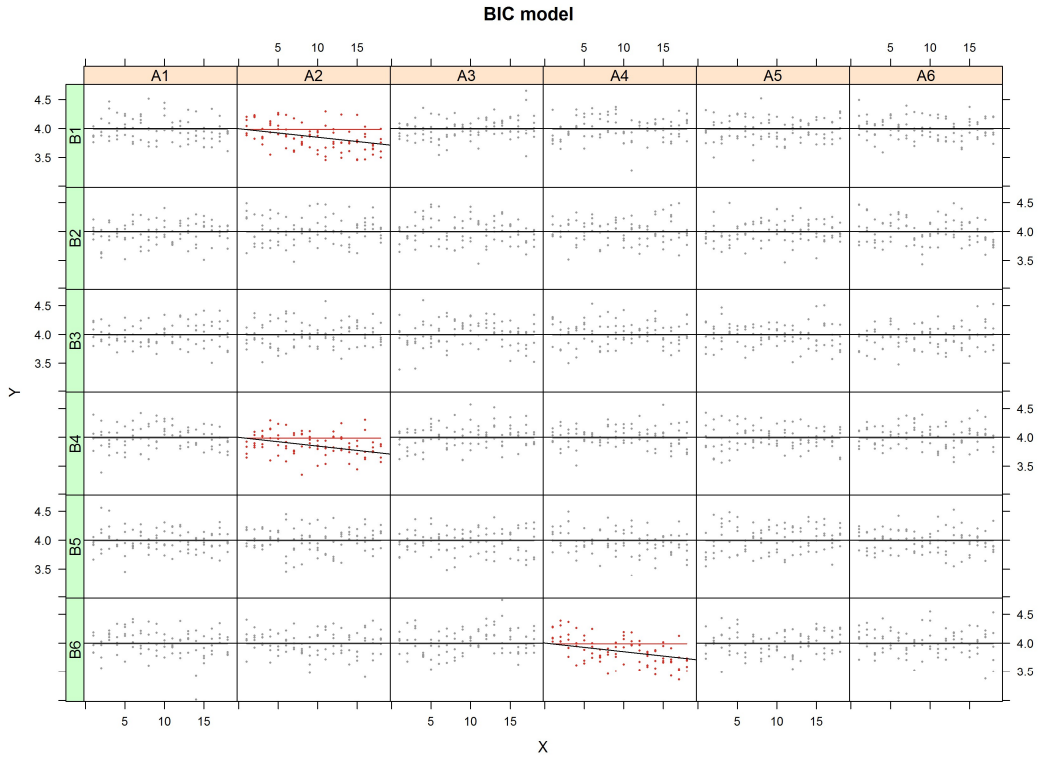
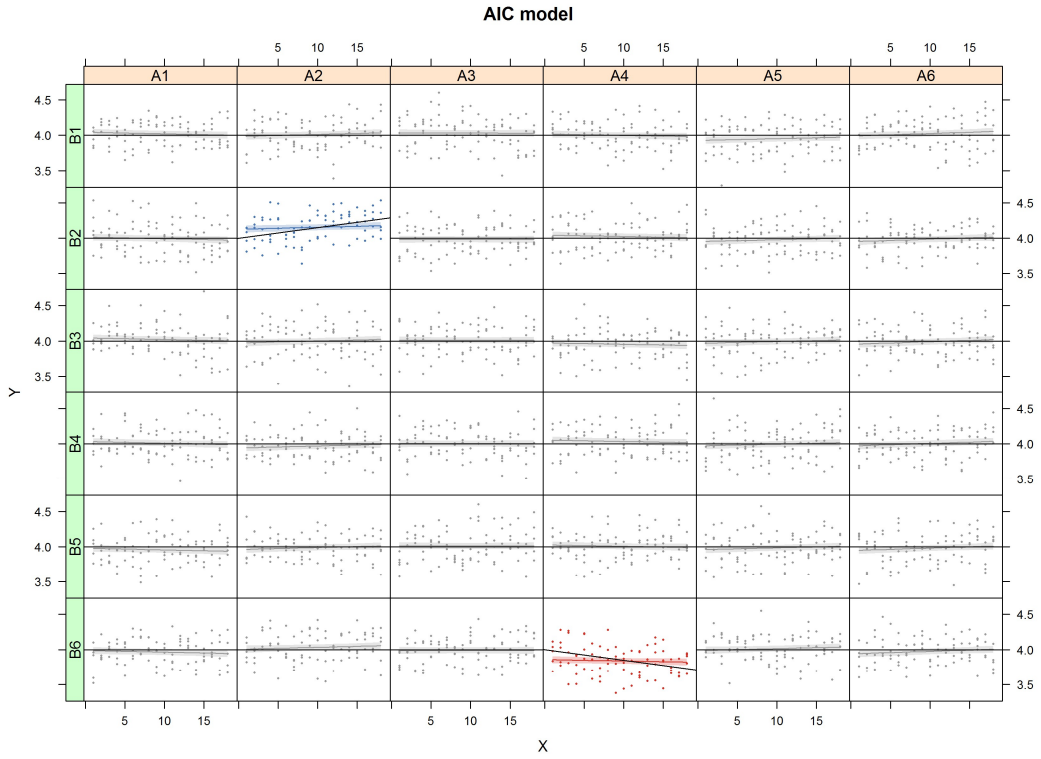


Figure B1. Model Type I: no trend effect. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.

a.



b.

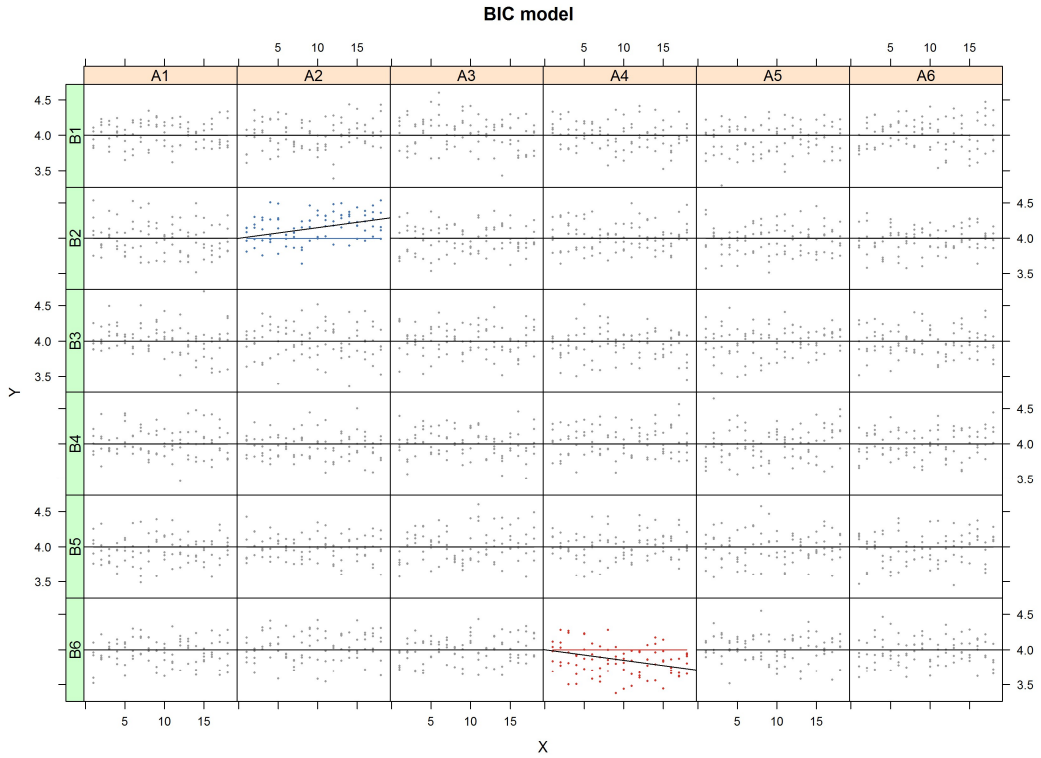
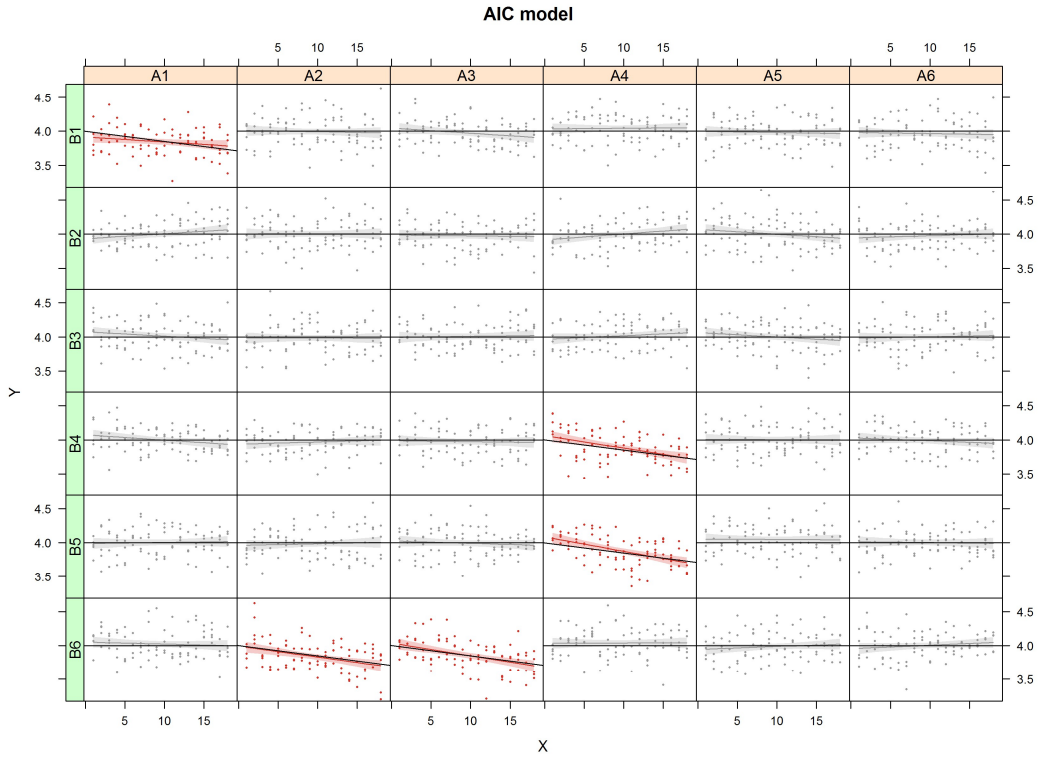


Figure B2. Model Type II: shift of intercept. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.

a.



b.

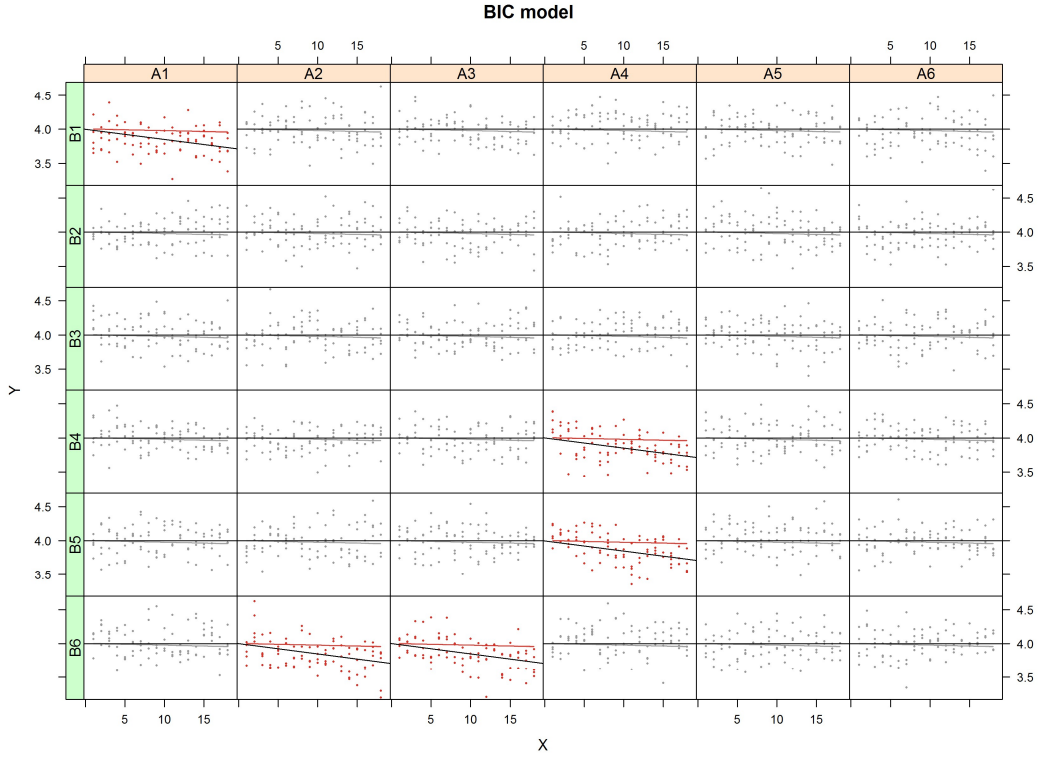
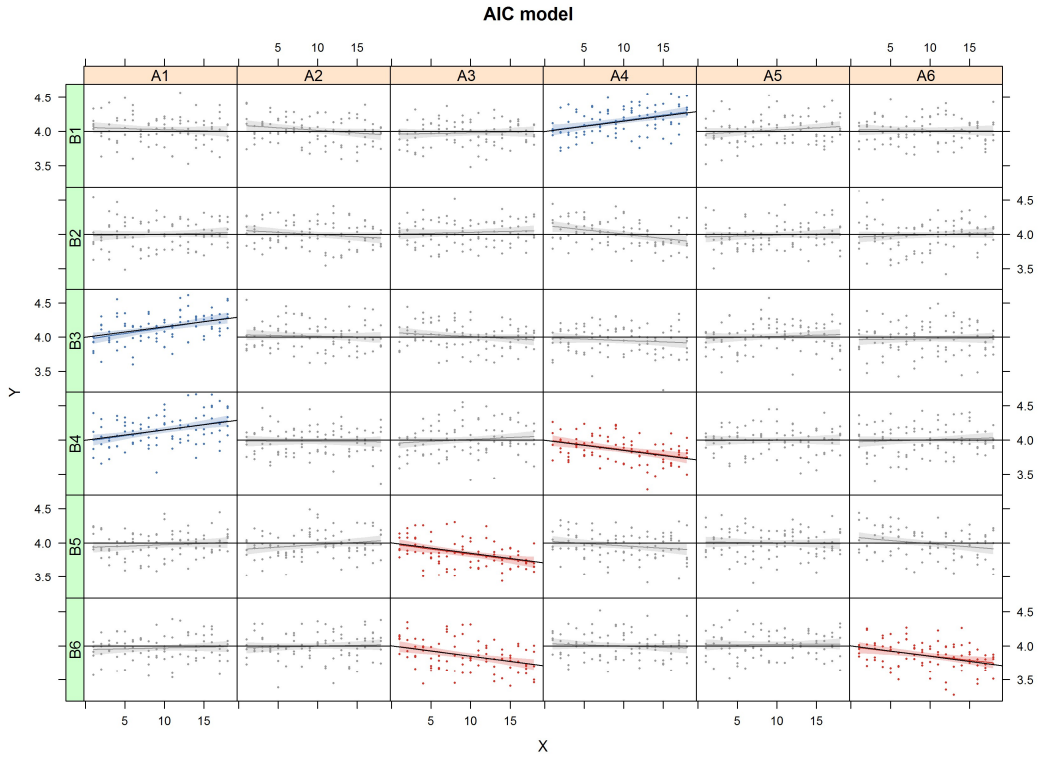


Figure B3. Model Type III: fixing slope. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.

a.



b.

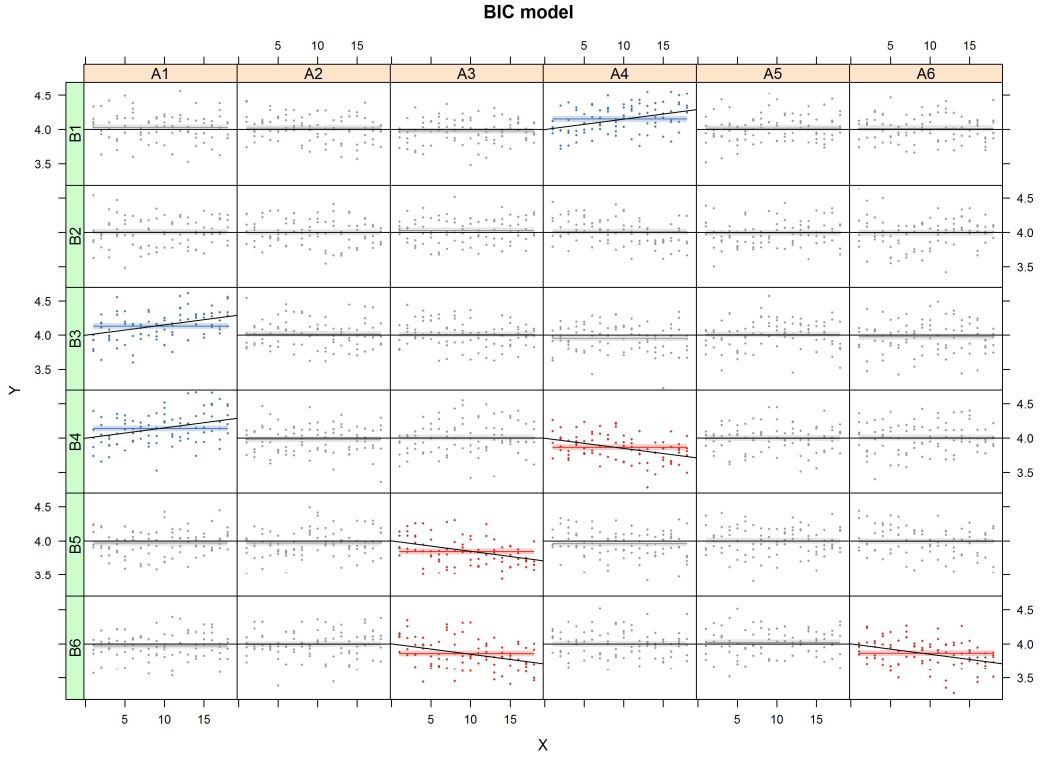


Figure B4. Model Type IV: row and column constraints. Fitted models for: a) AIC and b) BIC, showing simulated data and model fit (mean ± 95% CI) using colour to show whether data were simulated to have no trend (grey), negative trend (red) or positive trend (blue) in a cell. A solid black line is used to indicate the true data generating model in each cell.