Further clarification of the biases in and interpretation of regressions where catch is a predictor of penguin response

By

M.O. Bergh, OLRAC SPS
Silvermine House
Steenberg Office Park
Tokai 7945, Cape Town
South Africa

November 2014

Summary

This document explores the estimation characteristics of linear models which relate local pelagic biomass and local pelagic catches to island penguin response. The estimation problem is part of a well-documented class of linear and/ GLM regression problems that arise when there is predictor measurement error in the observed, and correlation between the unobserved true predictors (see for example Carroll et al, 1995 and Hardin et al 2003). Problems of bias are a known feature of such estimators.

A general statement of bias in estimates is presented here for models which do not incorporate the island or year categorical variables employed in Robinson (2013) (i.e. for the single island situation as per Appendix A). A two island extension of this analysis (Appendix B) suggests that the Appendix A single island biases are very similar for Robinson's GLMs which analyze two islands simultaneously using year or a common pelagic biomass as a covariate (Appendix B).

Results in Appendices A and B show that when local biomass is measured with error or when it is replaced by an imperfect proxy such as total resource biomass in linear models linking penguin biological response to pelagic catch and pelagic biomass, and when the underlying correlation between local pelagic catch and true local biomass is positive, then (a) the estimate of the increase in penguin response due to an increase in pelagic catch is positively biased, and (b) the estimate of the impact of local pelagic biomass on the penguin response is negatively biased. These biases are potentially substantial.

Using data from a system in which fishing takes place unhindered by scientific dictates to predict what would happen when fishing is directed by experimental constraints requires considerable care, even were the local pelagic biomass estimates available. Aside from correcting for the positive bias alluded to above, such an exercise requires that the positive impact that pelagic biomass has on catch is isolated from the negative impact that catch has on pelagic biomass. Predictions of the catch = zero situation should be based on a version of the model in which the positive impact of pelagic biomass on catch is absent. Failure to make this correction results in a potential third kind of bias in predicting the catch = zero situation for penguins, (c), where (a) and (b) above are the other two.

A model which has the necessary complexity to understand all these effects, if linear, is an unidentified non-recursive structural equation model which is described here. Although this SEM is strictly unidentified, identification can be achieved if the impact of catch on biomass can be set ***a priori***, perhaps using a reasoned mathematical argument.

There are two options for estimating penguin response to island closure. One option is to correct the Robinson (2013) method for bias, a suggested approach is outlined briefly in Appendix C. The other option is to make use of the flag variable indicating whether an island was open or closed to fishing during the experimental period in GLMM analyses, since this estimates the catch = zero situation directly.

Background to SEMs

The following is intended to clarify the arguments in MARAM/IWS/DEC14/PENG/A2 in relation to the use of catch as a covariate in GLMs for studying penguin response to experimental island closures. These arguments are constructed on the basis of linear models and relationships. If non-linear relationships are involved in the underlying mechanisms for the relevant range of values for variables of interest (local catch, biomass and penguin response data), then the logical implications are considerably more diverse than is described here.

The most general representation of the system considered here for a single island is the non-recursive SEM represented in Figure 5.

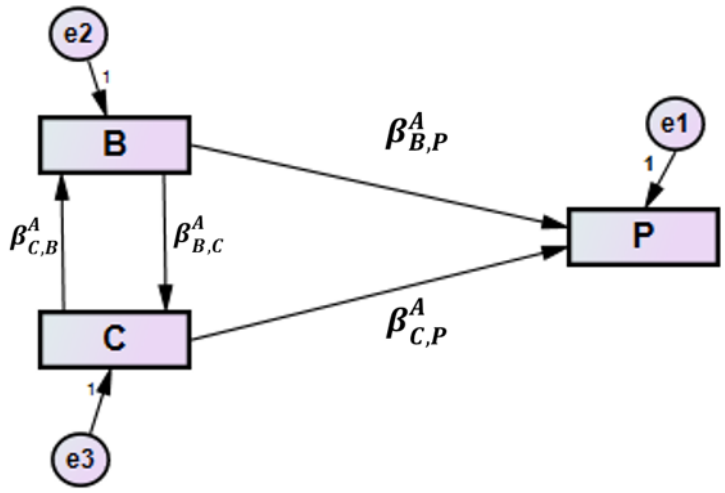

Figure 5. Model A: A path diagram representing the situation in which catch influences biomass, biomass influences catch, and both catch and biomass have a direct influence on penguin response.

The notation for Fig. 5 is:

B; local pelagic biomass, C; local pelagic catch, P; penguin response at island, $\beta^A_{B,P}$; the standardized regression weight for the linear impact of B on P for Model A, $\beta^A_{C,P}$; the standardized regression weight for the linear impact of C on P for Model A, $\beta^A_{C,B}$; the standardized regression weight for the linear impact of C on B for Model A, $\beta^A_{B,C}$; the standardized regression weight for the linear impact of B on C for Model A.

The symbol MB is used at a later stage to represent the resource wide pelagic biomass which has been used in previous GLMs in the absence of estimates of the local abundance estimate B. The following statements characterize the scope of the relationships between B, P and C with regard to their sign and not their absolute scale:

- The C to B causality is most likely negative since by definition catch reduces biomass.
- The B to C causality is most likely positive since pelagic fishing is an imperfect foraging system in which larger local biomasses tend on average to cause larger local catches.
- The B to P causality is most likely positive due to the fact that more food is good for penguins.
- The C to P causality is ambiguous. A positive C to P causality could be due to shoal dispersal, a negative C to P causality could be due to vessel avoidance.

Implications of scientific experimentation with the catch: When catch is experimentally manipulated then the B to C causal relationship is broken, i.e. the foraging incentive which causes larger catches when biomass is larger is switched off. This situation is represented by Figure 6. By closing and then opening islands to fishing, the mechanism governing the relationship between C, B and P is switching between Figure 5 (Island Open) and Figure 6 (Island Closed).

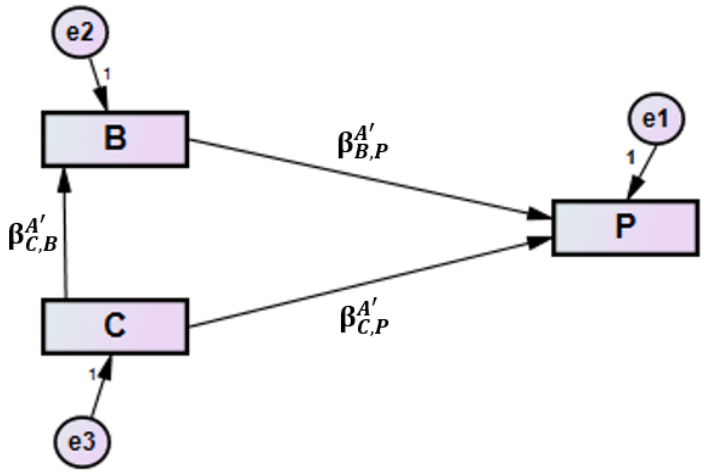

Figure 6. Model A$^{/}$, as for Model A, but in which the causal relationship from B to C has been removed (or set to zero). This represents the Island Closed status for the closure experiment.

OLS (Ordinary Least Squares) regression as employed in Robinson (2013) is represented by Figure 7.

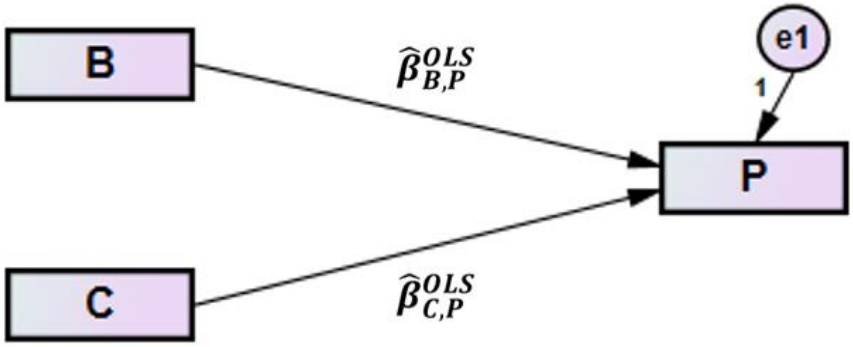

Figure 7. Model OLS. A path diagram representing Ordinary Least Squares regression (OLS) which is analogous to the GLMs carried out in Robinson (2013), except that this diagram does not represent any of the categorical variables in Robinson's (2013) GLMs. This diagram is representative of the one island situation only.

Two questions are posed here given this context and in relation to the island closure experiment:

Question 1: Does a Fig. 7 OLS regression provide an unbiased basis for predicting outcomes for C=0 when Fig. 5 is in operation?

Appendix A shows that an OLS regression produces biased estimates of the scale of linear relationships when the underlying process is a non-recursive SEM such as Figure 5 (although for simplicity Appendix A replaces the loop in Fig. 5 by a covariance relationship with no impact on the final results or their relevance to Fig. 5), and when there is measurement error in quantities being used in the regression analysis. Appendix B indicates that the full Type II GLM of Robinson's (2013) analysis is subject to a very similar level of biomass.

Consider as one illustration the situation where the C to P relationship is zero, but there is a net positive correlation between B and C, and B to P is a positive relationship. If B is measured without error, then the regression will correctly assign effects between C and B. In this case there is a spurious correlation between C and P because of the positive correlations between B to P and B to C. This causes problems when the value of B is not known reliably, and a proxy variable such as a resource wide estimate of biomass is used in its stead, say MB. There will be substantial lack of agreement between the true local pelagic biomass B and the resource wide figure MB. Under these circumstances there will be residual variance in P which MB cannot explain, but which C will be able to explain as a result of its correlation to B. This correlation results in an estimated positive relationship between C and P, even though none exists, as demonstrated in Appendix A.

Appendix A and B also demonstrates substantial negative bias in the estimate of the regression weight relating biomass to penguin response, when the biomass used contains measurement error. This bias is most likely predominantly the typical attenuation bias that arises when a predictor in linear regression is contaminated by measurement error. It is nonetheless a serious issue for the estimation of the effect of closure on penguin response, and could give rise to model selection errors and errors in testing the hypothesis that the biomass penguin link is zero. A number of methods exist and can be explored to correct for this attenuation bias (see Hardin et al 2003). Appendix C also explores a SEM based bias correction procedure.

The general answer to Question 1 is that if there is correlation between local catch and biomass, albeit weak, then using the OLS regression approach to predict the catch = zero situation produces substantively biased regression weight estimates.

Question 2: Does a Fig 5 SEM provide a basis for predicting outcomes for C = 0 when Fig 6 is in operation? If not what other options are available.

If it were possible to resolve the C to B and B to C regression weights then the answer to this question would be a cautious yes. However, Fig. 5 is an unidentified SEM for which the degrees of freedom are negative 1, and so this model cannot be fitted, and it is therefore not in principle possible to disentangle the scale of the negative C to B relationship from the positive B to C relationship without further information. This exercise is further complicated when B is replaced by MB, since this biases both the estimate of the impact of catches on penguin response, and the biomass to penguin regression weight estimate.

There are thus two main options:

Option 1: This option involves accumulating data for the system when the positive impact of B on C is switched off, i.e. catch is manipulated purely experimentally over time (Fig. 6 system in operation). This does not seem to be a meaningful option given the costs and time frames involved. Collecting data to characterize the behavior of the system when catches are determined experimentally, and using a model fit to these data to predict the catch = zero situation when an experiment has in fact been run with catch = zero is somewhat circuitous.

Option 2: A second option is to recognize that the impact of catch on biomass in Fig. 5 may be inferred heuristically. Catch reduces biomass by an amount that is related to the amount caught. In practice the impact is more complex depending on the exact nature of the system. It may nevertheless be possible to logically deduce the regression weight for the relationship C to B, say α. Inferring a value for this factor at, say α, as in for e.g. Fig. 9, would 'identify' the SEM depicted in Fig. 5. Given sufficient data on C, B and P at the local island level, the model could then be fitted. Predictions of the C=0 situation would then be obtained by using an adjunct to Fig. 9 where the B to C link is absent, viz. Fig. 10.

The practical application of this option is limited by the non-availability of local pelagic biomass estimates. Use of MB instead, the resource wide biomass estimate, introduces substantial bias into predictions for catch = zero. It appears from Appendix A that some of the positive impact of biomass on penguin response is lost from the B to P regression weight (which suggests that less catch and hence more biomass is good for penguins) and some non-existent positive impact is gained by the estimated C to P regression weight (which suggests that less catch is bad for penguins) – under the C to B positive correlation scenario of course. ***Appendix B extends this result to the two island situation and hence to the Type II GLMs actually used by Robinson (2013) and shows that most likely the bias in this case is very similar to the single island analysis results presented in Appendix A.***

The safest option in the context of this approach (i.e. assuming there was no closure experiment) thus seems to be:

(1) to correct the estimated C to P regression weight for bias (using methods such as in Hardin et al 2003, or the suggestion here in Appendix C),
(2) to correct the estimated biomass to penguin regression weight for bias (using methods such as in Hardin et al 2003, or Appendix C).

(3) to use some reasoning to predict how much local biomass will increase by if catch is set to zero,
(4) to use the Fig. 10 SEM model to estimate the catch=zero response for penguins.

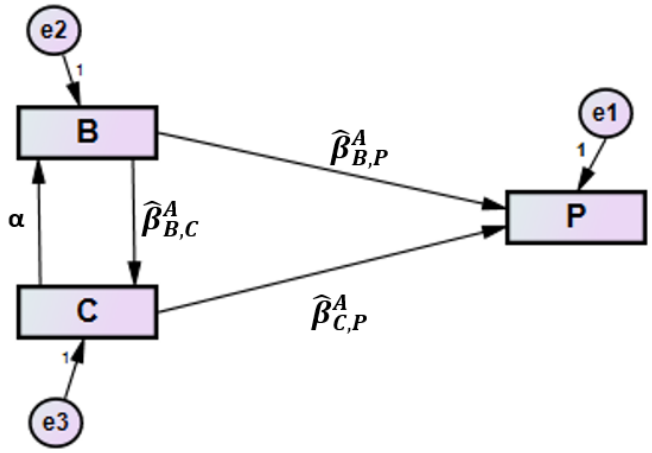

Figure 9. A version of Model A in which the causal relationship from C to B has been set to α based on mathematical reasoning.

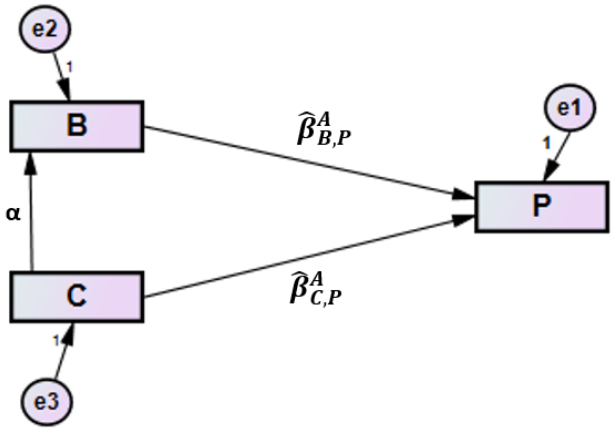

Figure 10. A version of Model A in which the causal relationship from C to B has been set to α based on mathematical reasoning, and the causal effect from B to C has been set to zero to model the impact of closing the area around an island to pelagic fishing.

Concluding Remarks

In light of the above, the following statement in MARAM/IWS/DEC14/Peng/B10 is revisited:

"…Figure 3 … shows the time series of sardine and anchovy catches made within different distances (and particularly within the sometimes closed area within 10 nm) of these islands. What is immediately apparent is that catches when this area when open span a wide range, including some very some years of very small catches. It would seem to make little sense to assume that the possible effect of these very small catches on penguin reproductive success is the same as that of much larger catches, but quite different to that in the absence of any catches."

This point touches on the heart of the matter that this paper is designed to address. The comment is advocating to use catch as a basis of predicting the implications of closure for penguins, citing instances when catch was very low, implying that this is tantamount to sampling a closed island situation. This statement is only true if the reason for the catch being so low is entirely unrelated to local abundance levels.

If on the other hand local catches are on average low only when local abundance levels are low, then there is really no idea as to what would happen under average local biomass conditions when catches are kept low by deliberate closures. If these catches being low do tend to coincide with low levels of local abundance, then our impression of what closures would do for penguins is only for times when local abundance is poor, so obviously we would then tend to conclude that closures are bad for penguins.

Given this problem, there is the temptation to introduce biomass as a covariate to remove this complication. The results presented here show that such an approach is unbiased if the biomass covariate that is used measures local abundance without error. However, if this measure of biomass is an imperfect measure of local abundance, then serious biases arise, even at a local biomass to catch correlation of as low as 0.200:

- The first of these biases is that where there is no direct relationship between catch and penguins, a positive relationship will emerge.
- The second is that the estimate of the importance of biomass for penguins is negatively biased.
- Thirdly, in order to use this schema to predict what closure would do, one has to account for the fact that a reduction in catch causes a relative increase in biomass, with a positive knock on effect for penguins. Considerable care is required to implement this correctly.

This approach only therefore seems reasonable if all the biases can be ironed out of the method and if the knock on effect mentioned is correctly calculated. Approaches such as are reported in Robinson (2013), i.e. to determine the impact of closure by setting the catch to zero in the regression equation, do not address the knock on effect or correct for the inherent biases highlighted in Appendix A. There may nevertheless be some potential to revise the Robinson (2013) method to address the problems of bias and other issues raised here (as per Hardin et al 2003 or Appendix C).

The use of a revised Robinson (2013) approach does however have to be weighed up against the straightforward alternative of using GLMs based on island, island closure, year (and month and even chick/nest) and pelagic biomass as predictors, in which the island closure effect is more simply estimated and the bias due to measurement error is not a problem. Although the view here is that this is a preferable approach, it does appear that there is a need for additional years of experimentation to provide sufficient power for such an approach. Additional experimentation may also be required for a revised Robinson (2013) approach, an assessment of which is beyond the scope of this document.

The estimation problems dealt with here are well known problems of estimation when there is measurement error in and correlation between predictors in linear regression and / or GLMs. Some points to note are (see for e.g. Hardin et al, 2003 or Carroll et al, 1995):

- The interpretation of the size of the regression for a particular independent variable can be misleading (depending on the scale and sign of correlations between independent variables, and the scale of measurement error),
- The analyst is required to attempt to correct for bias before drawing conclusions about affect sign or size (Appendix C provides one example of a bias correction approach).

Application of this advice to the specific penguin situation is that:

- The interpretation of the influence of catch on the dependent variable penguin response prior to any bias correction is inappropriate – similarly the OLS based estimates of the significance of the impact of pelagic biomass and catch on penguin response is biased, and therefore the model selection process is confused and misled,
- Predictions of what might occur under a new situation (e.g. island closure) require that biases in estimates are addressed.

Simulations could play a valuable role in determining the appropriate bias corrections applicable (see e.g. Appendix C). It is noted that the international panel has previously recommended that these simulations be carried out, but due to time/resource constraints this has not been possible.

References

Carroll, R. J., Ruppert, D., and Stefanski, L. A. 1995. Measurement error in non-linear models. New York, Wiley.

Fuller, W. A. 1987. Measurement Error Models. New York, Wiley.

Hardin, J.W., Schmiediche, H. and J.C. Raymond. 2003. The regression-calibration method for fitting generalized linear models with additive measurement error. The Stata Journal (2003) 3, Number 4, pp. 361–372

Robinson, W. M. L. 2013. Modelling the impact of the South African small pelagic fishery on African penguin dynamics. PhD thesis, University of Cape Town. xiv + 207 pp.

Appendix A: Determination of the bias in the estimate of the impact of pelagic catch and biomass on penguin response, when the OLS uses an imperfect measure of local pelagic abundance.

This exercise works directly off the correlation matrices linking the following variables:

- **C** – the local pelagic catch
- **P** – the penguin response
- **B** – the local pelagic biomass
- **MB** – an error prone proxy for B

A correct SEM (Structural Equation Model) describing this situation, Model C, is as follows:

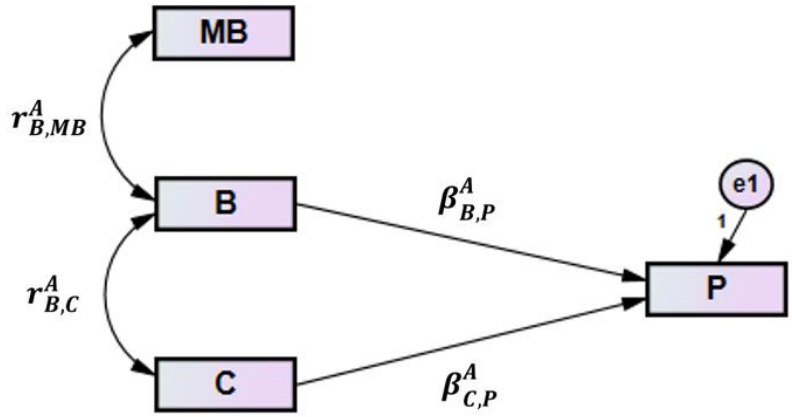

Figure A1. Model C: A representation of the interrelationship between catch (C), true biomass (B), measured or proxy biomass (MB), and penguin response (P) which captures the scope of linear interrelationships between these variables in the ongoing debate. Although biomass is shown here as an observed quantity it is in fact not available, and the only available measure of biomass is "MB" which is "B" contaminated by measurement error. This diagram does not explicitly represent the covariance between C and MB and between P and MB, these quantities are implied, and no more or less covariance than the implied amount is assumed here, without loss of applicability of the final results. Model C is equivalent to Model A w.r.t to the relationship between C, P and B, the only difference being the introduction of MB and the replacement of the loop from C to B and back from B to C by the covariance between B and C. Hence the subscript denoting the model version is shown as 'A' and not 'C'.

Associated with Model C is its implied correlation matrix, which can be inferred from the diagram in Figure A1 and Wright's Laws (see below). This is shown in Fig. A2.

	MB	B	C	P
MB	1			
B	$r^A_{B,MB}$	1		
C	$r^A_{B,C}r^A_{B,MB}$	$r^A_{B,C}$	1	
P	$\beta^A_{B,P}r^A_{B,MB}$	$\beta^A_{B,P}+\beta^A_{C,P}r^A_{B,C}$	$\beta^A_{C,P}+r^A_{B,C}\,\beta^A_{B,P}$	1

Figure A2. The implied correlation matrix associated with the path diagram in Fig. A1. This follows from the application of Wright's Laws to the path diagram for Model C.

If Model C is absolutely correct, then the implied correlation matrix in Fig. A2 is identical to the sample correlation matrix for the underlying data available for fitting the model. That is, a perfect fit between Model C and the data available to fit the model is being assumed. Consequently the matrix in Fig. A2 can be taken as the sample correlation matrix representing the relationship between P, C, B and MB. (Note: Fitting any of the SEMs shown here is equivalent to obtaining the best fit between the implied and sample covariance and correlation matrices. OLS regression can be cast as a fit of this type.)

Because in reality the true values of B are not available and only MB can be used, the applicable SEM which is implicitly being fitted by for e.g. Robinson (2013) using OLS regression is the path diagram in Figure A3, a variant of Model C in which B and MB have been interchanged. In this path diagram the value of B is represented in order to complete the implied correlation matrix, however none of what follows assumes any knowledge of B ***other than the assumption of the B to P relationship and the correlation between B and MB.***

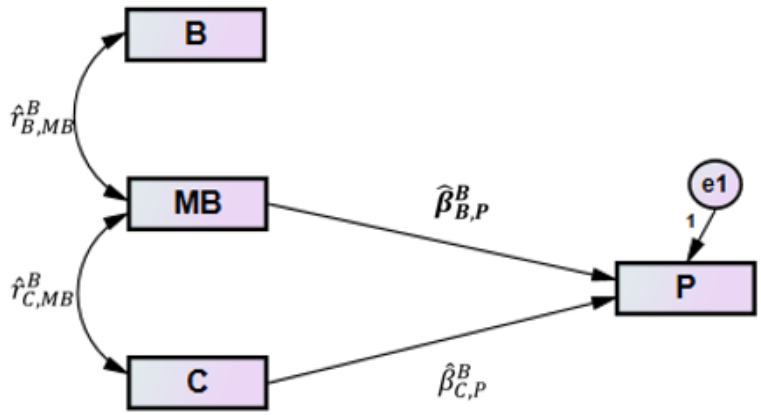

Figure A3. Model B: This is effectively Model C (or equivalently Model A) in which the positions of B and MB have been interchanged. Aside from the inclusion of B, this is also a very close representation of the

Ordinary Least Squares equation used by Robinson (2013) for certain GLMs where P is the target variable and MB and C are predictors, for the single island situation, except that it includes the double headed arrow representing the covariance between C and MB. Inclusion of this covariance term leads to results which differ by a small amount compared to the results obtained from Model OLS in Fig. 7. Model B is used instead of Model OLS because some simple algebraic results were found which allowed results to be quickly characterized. In this model $\widehat{\beta}^{B}_{B,P}$ should strictly be denoted $\widehat{\beta}^{B}_{MB,P}$, however the former representation is retained since it is regarded as the best estimate of the latter.

Model B in Fig. A3 has the implied correlation matrix shown in Fig. A4.

	MB	B	C	P
MB	1			
B	$\hat{r}^{B}_{B,MB}$	1		
C	$\hat{r}^{B}_{C,MB}$	$\hat{r}^{B}_{C,B}$	1	
P	$\widehat{\beta}^{B}_{B,P}+\widehat{\beta}^{B}_{C,P}\hat{r}^{B}_{C,MB}$	$\hat{\beta}^{B}_{B,P}\hat{r}^{B}_{B,MB}$	$\widehat{\beta}^{B}_{C,P}+\hat{r}^{B}_{C,MB}\widehat{\beta}^{B}_{B,P}$	1

Figure A4. The implied correlation matrix associated with the path diagram shown in Fig. A3, Model B. As before, these follow from the application of Wright's Laws.

The parameters of Model B can be determined by fitting the implied correlation matrix in Fig. A4 to the underlying sample correlation matrix (in practice most SEMs are fitted to the unstandardized sample covariance relationship, but the net result is tantamount to a fit of the standardized quantities). As argued above, and since Model C is the correct representation of the system and not Model B, the applicable sample correlation matrix is given by Fig. A2. The parameters for this fit are all the quantities in Fig A4 which are accented. ***If it is known that the fit of Model B to the implied correlation matrix is a perfect fit then the fit can be carried out by equating cells in the Model B correlation matrix with the cells in the sample correlation matrix and solving for parameters of interest, rather than needing to resort to SEM fitting software. It can however be shown that the fit in question is <u>not perfect</u> since there is 1 degree of freedom. It can be rendered perfect by including as a model parameter the correlation between B and e1 – this leads to a model with zero degrees of freedom. Inclusion of this change impacts only the implied correlation in the (B,P) cell. Thus by assuming that the fit for Model B as written is perfect, one must discount the (B,P) cell in what follows.*** The fit/solution can be achieved in two steps. In the first step a number of accented quantities can be replaced by known quantities from Fig. A2, as follows:

	MB	B	C	P
MB	1			
B	$r^{A}_{B,MB}$	1		
C	$r^{A}_{B,C}r^{A}_{B,MB}$	$r^{A}_{B,C}$	1	
P	$\hat{\beta}^{B}_{B,P}+\hat{\beta}^{B}_{C,P}r^{A}_{B,C}r^{A}_{B,MB}$	$\hat{\beta}^{B}_{B,P}r^{A}_{B,MB}$	$\hat{\beta}^{B}_{C,P}+r^{A}_{B,C}r^{A}_{B,MB}\hat{\beta}^{B}_{B,P}$	1

Figure A5. A first step simplification of the implied correlation matrix in Fig. A4 which occurs when this is equated to the sample correlation matrix in Fig. A2.

The next step in fitting the matrix in Fig. A5 to the sample correlation matrix in Fig. A2 is to equate the (P,MB) and (P,C) cells in each of these matrices. ***Although cell (P,B) also contains information about $\hat{\beta}^{B}_{B,P}$, it should not be included in the solution process - this is because we know that the model is perfect when the B to e1 correlation is included, and that this modification only impacts the (P,B) implied correlation and does not affect cells (P,MB) and (P,C)***. The following two equations are thus available for solving for the parameters of interest:

$$\hat{\beta}^{B}_{B,P}+\hat{\beta}^{B}_{C,P}r^{A}_{B,C}r^{A}_{B,MB}$$

$$\hat{\beta}^{B}_{C,P}+r^{A}_{B,C}r^{A}_{B,MB}\hat{\beta}^{B}_{B,P}$$

This is a two equation system involving, as parameters of interest, the best estimate of the influence of biomass on penguins $\hat{\beta}^{B}_{B,P}$ and of catch on penguins $\hat{\beta}^{B}_{C,P}$ in terms of the correlation between B and C, $r^{A}_{B,C}$, the correlation between B and MB, $r^{A}_{B,MB}$, the true impact of biomass on penguins $\beta^{A}_{B,P}$ and the true impact of pelagic catch on penguins $\beta^{A}_{C,P}$. The solution is as follows:

$$\hat{\beta}^{B}_{B,P}=\frac{(\beta^{A}_{B,P}r^{A}_{B,MB}-r^{A}_{B,C}r^{A}_{B,MB}(\beta^{A}_{C,P}+r^{A}_{B,C}\beta^{A}_{B,P}))}{(1-(r^{A}_{B,C}r^{A}_{B,MB})^{2})}$$

$$\hat{\beta}^{B}_{C,P}=\frac{(\beta^{A}_{C,P}+r^{A}_{B,C}\beta^{A}_{B,P}(1-r^{A}_{B,MB}{}^{2}))}{(1-(r^{A}_{B,C}r^{A}_{B,MB})^{2})}$$

The behavior of the estimators $\hat{\beta}^{B}_{B,P}$ and $\hat{\beta}^{B}_{C,P}$ can be explored for selected values of $r^{A}_{B,C}$, $r^{A}_{B,MB}$, $\beta^{A}_{B,P}$ and $\beta^{A}_{C,P}$, as is reported below in Tables A1 –A4 and Figures A6 – A8:

Table A1. The estimates of $\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{B,P}}$ and $\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{C,P}}$, for $\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{C,P}} = \mathbf{0}$, and for some selected values of $\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,C}}$, $\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,MB}}$ and $\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{B,P}}$ subject to the constraint $\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,C}} > \mathbf{0}$. These results were actually achieved using SEM software to fit Model B to the sample correlation matrix – these results agree with the formulaic solutions.

$\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,MB}}$	$\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,C}}$	$\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{B,P}}$	$\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{C,P}}$	$\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{C,P}}$	$\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{B,P}}$
0.400	0.300	0.800	0.00	0.205	0.295
0.600	0.300	0.800	0.00	0.159	0.451
0.800	0.300	0.800	0.00	0.092	0.618
0.600	**0.200**	0.800	0.00	0.104	0.468
0.800	**0.500**	0.800	0.00	0.171	0.571

These results not only show the potential for positive bias in the estimates of the impact of local pelagic catch on local pelagic biomass, they also show substantial negative bias in the estimates of the impact of local pelagic biomass on penguin response.

Table A2. The estimates of $\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{B,P}}$ and $\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{C,P}}$, for $\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{C,P}} > \mathbf{0}$, and for some selected values of $\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,C}}$, $\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,MB}}$ and $\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{B,P}}$ subject to the constraint $\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,C}} > \mathbf{0}$. These results were actually achieved using SEM software to fit Model B to the sample correlation matrix – these results agree with the formulaic solutions.

$\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,MB}}$	$\boldsymbol{r}^{\boldsymbol{A}}_{\boldsymbol{B,C}}$	$\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{B,P}}$	$\boldsymbol{\beta}^{\boldsymbol{A}}_{\boldsymbol{C,P}}$	$\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{C,P}}$	$\widehat{\boldsymbol{\beta}}^{\boldsymbol{B}}_{\boldsymbol{B,P}}$
0.400	0.300	0.800	0.28	0.489	0.261
0.600	0.300	0.800	0.28	0.448	0.399
0.800	0.300	0.800	0.28	0.389	0.547
0.800	**0.500**	0.800	0.28	0.505	0.438
0.600	**0.200**	0.800	0.28	0.388	0.433

Table A3. The estimates of $\widehat{\boldsymbol{\beta}}_{B,P}^{B}$ and $\widehat{\boldsymbol{\beta}}_{C,P}^{B}$, for $\boldsymbol{\beta}_{C,P}^{A} < 0$, and for some selected values of $\boldsymbol{r}_{B,C}^{A}$, $\boldsymbol{r}_{B,MB}^{A}$ and $\boldsymbol{\beta}_{B,P}^{A}$ subject to the constraint $\boldsymbol{r}_{B,C}^{A} > 0$. These results were actually achieved using SEM software to fit Model B to the sample correlation matrix – these results agree with the formulaic solutions.

$\boldsymbol{r}_{B,MB}^{A}$	$\boldsymbol{r}_{B,C}^{A}$	$\boldsymbol{\beta}_{B,P}^{A}$	$\boldsymbol{\beta}_{C,P}^{A}$	$\widehat{\boldsymbol{\beta}}_{C,P}^{B}$	$\widehat{\boldsymbol{\beta}}_{B,P}^{B}$
0.400	0.300	0.800	**-0.28**	-0.080	0.330
0.600	0.300	0.800	**-0.28**	-0.131	0.504
0.800	0.300	0.800	**-0.28**	-0.205	0.689
0.800	**0.500**	0.800	**-0.28**	-0.162	0.705
0.600	**0.200**	0.800	**-0.28**	-0.180	0.502

Table A4. The estimates of $\widehat{\boldsymbol{\beta}}_{B,P}^{B}$ and $\widehat{\boldsymbol{\beta}}_{C,P}^{B}$, for $\boldsymbol{\beta}_{C,P}^{A} > 0$, and for some selected values of $\boldsymbol{r}_{B,C}^{A}$, $\boldsymbol{r}_{B,MB}^{A}$ and $\boldsymbol{\beta}_{B,P}^{A}$ subject to the constraint $\boldsymbol{r}_{B,C}^{A} < 0$. These results were actually achieved using SEM software to fit Model B to the sample correlation matrix – these results agree with the formulaic solutions.

$\boldsymbol{r}_{B,MB}^{A}$	$\boldsymbol{r}_{B,C}^{A}$	$\boldsymbol{\beta}_{B,P}^{A}$	$\boldsymbol{\beta}_{C,P}^{A}$	$\widehat{\boldsymbol{\beta}}_{C,P}^{B}$	$\widehat{\boldsymbol{\beta}}_{B,P}^{B}$
0.400	-0.300	0.800	0.28	0.080	0.330
0.600	-0.300	0.800	0.28	0.131	0.504
0.800	-0.300	0.800	0.28	0.205	0.689
0.800	**-0.500**	0.800	0.28	0.162	0.705
0.600	**-0.200**	0.800	0.28	0.180	0.502

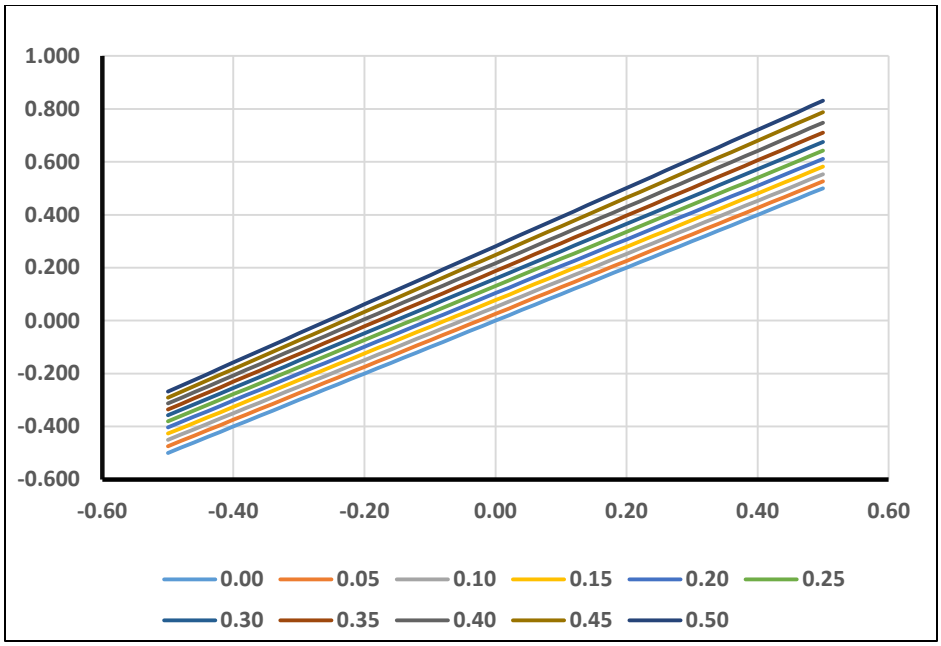

Figure A6. Estimates of $\widehat{\beta}^{B}_{C,P}$ (y-axis) for ranges of values of $r^{A}_{B,C}$ (different coloured lines), as a function of the true value $\beta^{A}_{C,P}$ (x-axis) for $r^{A}_{B,MB}$ = 0.6 and for $\beta^{A}_{B,P}$= 0.8. These results only use the formulaic solutions.

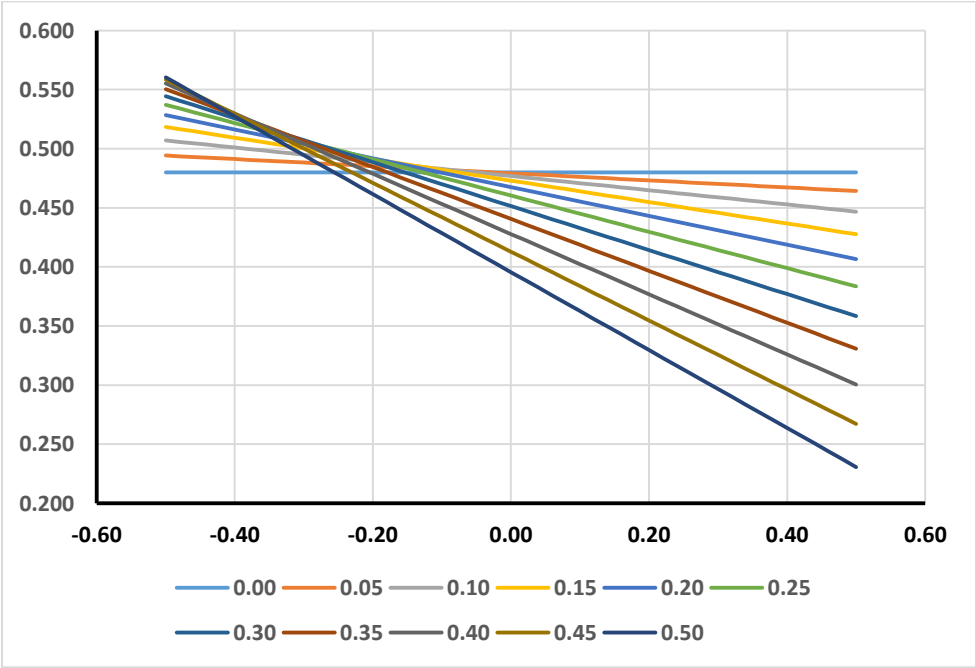

Figure A7. Estimates of $\widehat{\beta}^{B}_{B,P}$ (y-axis) for ranges of values of $r^{A}_{B,C}$ (different coloured lines), as a function of the true value $\beta^{A}_{C,P}$ (x-axis) for $r^{A}_{B,MB}$ = 0.6 and for $\beta^{A}_{B,P}$= 0.8. These results only use the formulaic solutions.

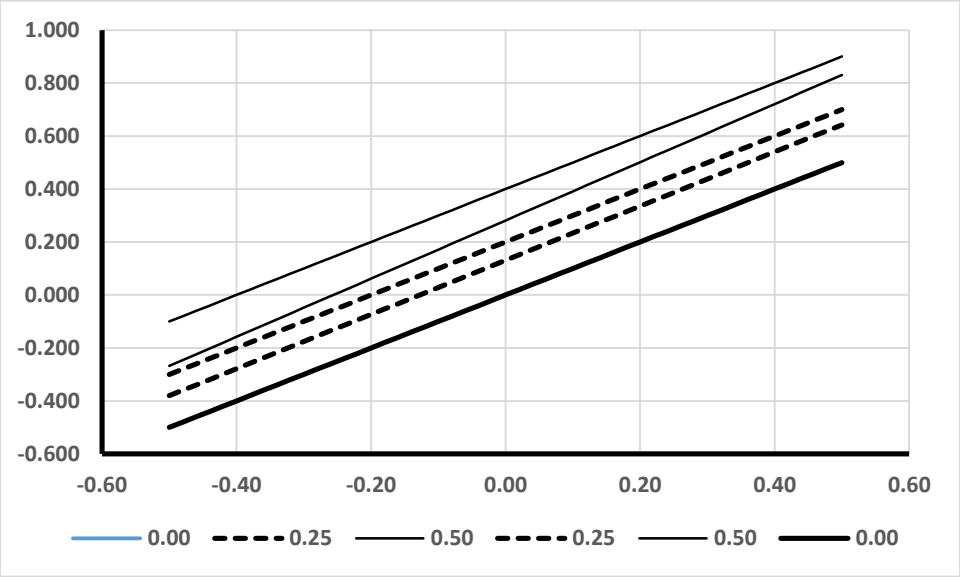

Figure A8. Estimates of $\widehat{\beta}^{B}_{C,P}$ (y-axis) for different values of $r^{A}_{B,C}$ (different lines), as a function of the true value $\beta^{A}_{C,P}$ (x-axis) for $r^{A}_{B,MB}$ = 0.6 and for $\beta^{A}_{B,P}$= 0.8, and when $\widehat{\beta}^{B}_{B,P}$ is either estimated (lower of the pair of lines) or $\widehat{\beta}^{B}_{B,P}$ is forced to zero (i.e. B/MB is excluded from the OLS) – upper of the pair of lines. These results show that when the B to C correlation is positive then the bias in the estimate $\widehat{\beta}^{B}_{C,P}$ is larger when $\widehat{\beta}^{B}_{B,P} = 0$, i.e. pelagic biomass is omitted from the OLS.

WRIGHT'S LAWS

Given a fitted SEM, Wright's Laws allow one to calculate the implied correlation matrix. Wright's Laws were used extensively here to establish the correct relationships between variables P, C, B and MB in Model A such that the fitted model and hence the implied correlation matrix conforms precisely with the sample correlation matrix. This procedure for revealing bias in the estimation process takes the sample correlation matrix as direct input, and thus avoids the need for lengthy Monte Carlo simulation studies. Note that correlation coefficients and beta coefficients are equivalent from the point of view of Wright's Laws

The implied correlation between V1 and V2 is the sum of the correlations from all permissible pathways linking these two variables. The definition of permissible pathways is as follows, where single headed arrows represent a directional causal relationship and double headed arrows represent covariance:

1. You can trace backwards along an arrow and then forwards but not the other way around.
2. You can only pass through a variable once in a pathway.
3. There may only be one two headed arrow in a pathway.

These laws are applicable to standardized estimates. The correlation contribution for a single pathway is obtained by multiplying all the r's and β's in that pathway. The total implied correlation between V1 and V2 is the sum of the correlation contributions from each pathway.

Loehlin, J.C. 2004. Latent variable models: An introduction to factor, path, and structural equation analysis, Lawrence Erlbaum.

Appendix B: Extensions to treat the two island situation, including a year effect or a common pool of pelagic biomass as a covariate

In order to extend the analyses to treat the two island situation, including a year effect or a common pool of pelagic biomass as a covariate, the following notation is used:

- **CD** – the local pelagic catch at Dassen Island
- **PD** – the penguin response at Dassen Island
- **BD** – the local pelagic biomass at Dassen Island
- **CR** – the local pelagic catch at Robben Island
- **PR** – the penguin response at Robben Island
- **BR** – the local pelagic biomass at Robben Island
- **MB** – a measure of B which is common to Dassen Island and Robben Island, which could alternatively be viewed as a year effect.

Given this notation, an applicable structural equation model is the model illustrated in Fig. B1 below, where initially it is assumed that the local pelagic biomass is a known quantity:

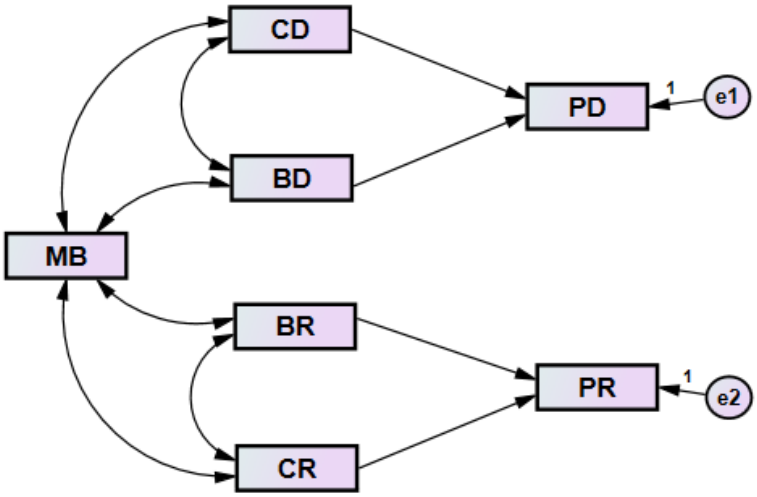

Figure B1. An SEM model, Model D, depicting the relationships that exists between biomass, catch and penguin response at two islands when there is a common measure of biomass that is correlated to the local pelagic biomass levels at each island. This model is also applicable to the situation where instead of common pelagic biomass, MB is exchanged for Y a year effect. It is suggested that this SEM is equivalent to the Type II GLMs employed in Robinson (2013).

The data underlying Model D has 28 sample moments, there are 17 free parameters and 11 degrees of freedom. Model E in Figure B2 is closely related to Model D, see below:

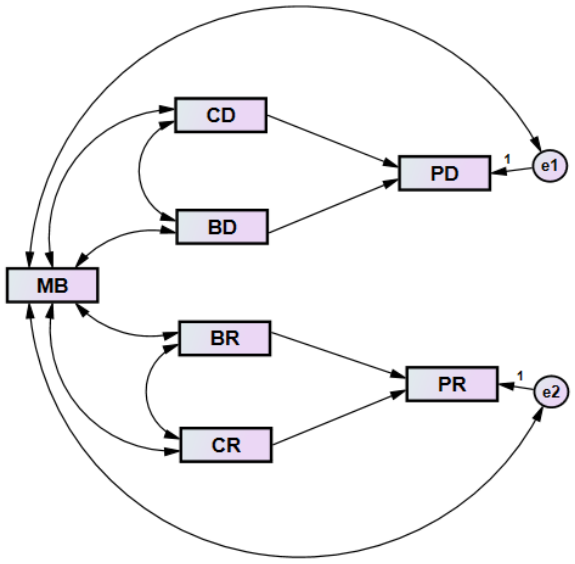

Figure B2. An SEM model, Model E, depicting the relationships that exists between biomass, catch and penguin response at two islands when there is a common measure of biomass that is correlated to the local pelagic biomass levels at each island. This model is a variant of Model D which includes explicit correlations between MB and the residuals e1 and e2.

Model E has 9 degrees of freedom. In order to explore potential bias in the OLS based GLM methods of Robinson (2013), the following example was used: $r_{BD,MB} = 0.400$, $r_{CD,BD} = 0.300$, $\beta_{BD,PD} = 0.800$, $\beta_{CD,PD} = 0.280$, $r_{BR,MB} = 0.600$, $r_{CR,BR} = 0.300$, $\beta_{BR,PR} = 0.800$, $\beta_{CR,PR} = 0.000$. Using these values, an implied correlation matrix for Model D was constructed setting all implied correlations between the two islands to zero, and applying Wright's Laws assuming no relationship between the two islands, since in reality none exist. This matrix is shown below as Fig. B3:

	MB	BD	CD	PD	BR	CR	PR
MB	1.000						
BD	0.400	1.000					
CD	0.120	0.300	1.000				
PD	0.320	0.884	0.520	1.000			
BR	0.600	0.000	0.000	0.000	1.000		
CR	0.180	0.000	0.000	0.000	0.300	1.000	
PR	0.480	0.000	0.000	0.000	0.800	0.240	1.000

Figure B3. The implied correlation matrix for Model D which is constructed when $r_{BD,MB} = 0.400$, $r_{CD,BD} = 0.300$, $\beta_{BD,PD} = 0.800$, $\beta_{CD,PD} = 0.280$, $r_{BR,MB} = 0.600$, $r_{CR,BR} = 0.300$, $\beta_{BR,PR} = 0.800$, $\beta_{CR,PR} = 0.000$.

Even though Model E has 9 degrees of freedom, it achieves a perfect fit to the implied correlation matrix in Fig. B3, implying that it is a correct representation of the relationships involved and also that the implied correlation matrix in Fig. B3 can be treated as a sample correlation matrix involving the 7 quantities under consideration.

Model Fa or b in Fig. B3 is a model in which the assumption is that the local pelagic biomass information is not available but rather a common pelagic biomass MB is used (relationships with BD and BR are nevertheless represented as if they were known, but not used in the OLS GLMs, with no impact on the final results or conclusions reached here):

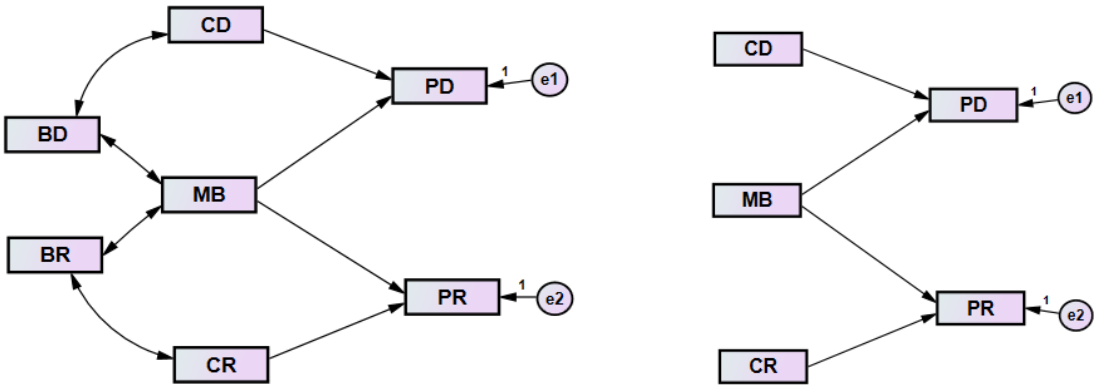

Figure B4. This is Model Fa,b, a simplified representation of the relationships between CD, BD, PD, MB, CR, BR and PR, intended to represent the two island GLM in Robinson (2013) where pelagic catch is used as a covariate together with either a common pelagic biomass or a year effect.

Model Fb is a representation of the relationships between CD, BD, PD, MB, CR, BR and PR, intended to represent the two island GLM in Robinson (2013) where pelagic catch is used as a covariate together with either a common pelagic biomass or a year effect. A further model relevant to this discussion is Model G shown below:

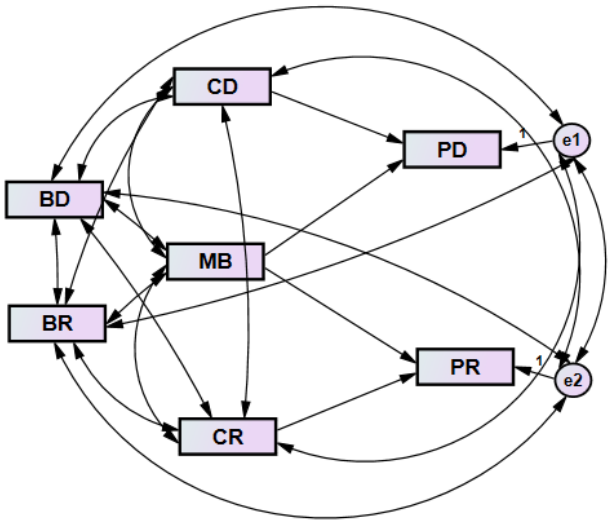

Figure B5. This is Model G, a variant of Model F, in which all relevant covariances are fitted. Model G consequently has zero degrees of freedom.

The numerical results shown in Table B1 were obtained (since an algebraic result for the SEM fit was not obtained at the time of writing, the IBM SEM software package AMOS (Analysis of Moment Structures) was used to carry out the fit):

Table B1. Estimates of $\beta_{CD,PD}$, $\beta_{BD,PD}$, $\beta_{BR,PR}$ and $\beta_{CR,PR}$ for Models E, Fa,b and G, as well as using Model B of Appendix A which carries out completely separate analyses for each island, denoted 'Separate Island Estimates'. Model E is the correct model so the estimates are equal to the true values. Model F is a close analogue of the GLMs in Robinson (2013) which use island as a categorical variable and both catch and biomass as covariates, or catch and a year effect as covariates (strictly year in this context is a fixed categorical effect). Model G is a variant of Model Fa which was found to reproduce the "Separate Island Estimates".

	True Value	Model E Estimates	Separate Island Estimates	Model Fa,b Estimates	Model G Estimates
$r_{BD,MB}$	0.400	omitted	omitted	omitted	Omitted
$r_{CD,BD}$	0.300	"	"	"	"
$r_{BR,MB}$	0.600	"	"	"	"
$r_{CR,BR}$	0.300	"	"	"	"
$\beta_{CD,PD}$	0.280	0.280	0.489	0.496	0.489
$\beta_{BD,PD}$ or $\beta_{MB,PD}$	0.800	0.800	0.261	0.265	0.261
$\beta_{BR,PR}$ or $\beta_{MB,PR}$	0.800	0.800	0.451	0.457	0.451
$\beta_{CR,PR}$	0.000	0.000	0.159	0.161	0.159

The results in Table B1 confirm that:

- Model E returns the true values as estimates, consistent with expectation (the model is correct).
- The "Separate Island Estimates", which treat every island separately using Model B of Appendix A, are identical to those produced using Model G.
- Model Fb, which is closest to the second type of GLMs employed in Robinson (2013), produces estimates which show slightly greater bias in the estimate of the catch – penguin regression than does Model B of Appendix A or Model G (which is a variant of Model F with zero degrees of freedom).

Appendix C: Exploration of bias correction options

A possible bias correction approach for the single island situation is to fit an SEM of the form shown in Fig. C1. This SEM uses an unobserved quantity BDTrue. This SEM is identified if three values are given. These are the variance of BDTrue, its correlation with CD and its correlation with MB. Table C1 illustrates preliminary results.

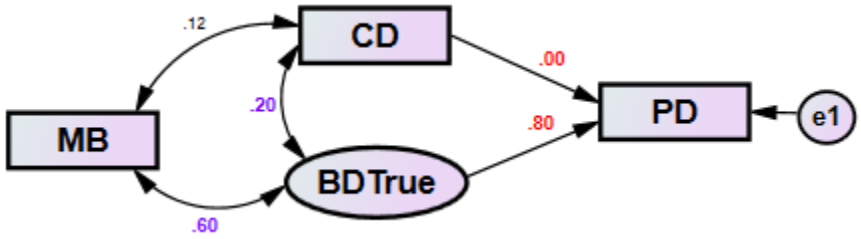

Figure C1. An SEM which was explored as a mechanism to achieve bias correction for the Type II GLMs used in Robinson (2013) (it is contended here that the GLM can be split into two SEMs for each island for these purposes). It involves incorporating an unmeasured quantity BDTrue for the true local pelagic biomass around Dassen Island (D in this case) in the SEM model. The SEM is identified if the variance of BDTrue, its correlation with CD and its correlation with MB is set. Thus, aside from the variance of BDTrue, two values need to be specified correctly to achieve a reliable bias correction.

The results of the attempted bias correction is shown below in Table C1. It appears that if the specification for the correlation between CD and BDTrue is incorrect then substantial biases persist. Other methods could be attempted, such as tuning the CD to BDTrue correlation until the CD to PD regression weight reaches zero (on the basis that given the arguments pro a positive and negative relationship, zero relationship is a suitable compromise) – exploration of this was out of scope here.

Table C1. The results of application of the bias correction procedure in Fig. C1.

	$r^A_{BD,MB}$	$r^A_{BD,CD}$	$\beta^A_{CD,PD}$	$\beta^A_{BD,PD}$	$\hat{\beta}^B_{CD,PD}$	$\hat{\beta}^B_{BD,PD}$	$\hat{\beta}^{Corrected}_{CD,PD}$	$\hat{\beta}^{Corrected}_{BD,PD}$
	0.600 (true)	**0.200 (true)**	0.00	0.800	0.104	0.468		
Bias Corrected	**0.600 (assumed)**	**0.200 (assumed)**	0.00	0.800	0.104	0.468	**0.00**	**0.800**
Bias Corrected	**0.700 (assumed)**	**0.150 (assumed)**	0.00	0.800	0.104	0.468	**0.059**	**0.676**
Bias Corrected	**0.900**	**0.00**	0.00	0.800	0.104	0.468	**0.160**	**0.512**
Bias Corrected	**0.400**	**0.00**	0.00	0.800	0.104	0.468	**0.160**	**1.152***

- Problems of negative variance (!) were encountered with the SEM fit (and correlations larger than 1!).