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.

(Revised 28 November 2014, to include the pathway from P to C to MB as a contributor to the implied correlation between P and MB in Fig A2 and following logic).

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:

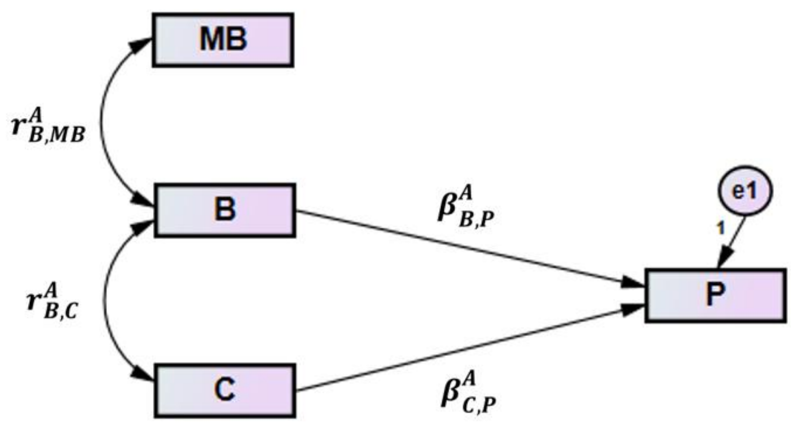

Figure A1. Model C: A representation of the interrelationship between catch (C), true biomass (B), measured or proxy biomass (MB), and penguin response (P) which captures the scope of linear interrelationships between these variables in the ongoing debate. Although biomass is shown here as an observed 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 e1 and MB, but these relationships are assumed to be present. 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), and assuming that the covariance from C to MB is accounted for in the model. This is shown in Fig. A2.

	MB	B	C	P
MB	**1**			
B	$\boldsymbol{r^A_{B,MB}}$	**1**		
C	$\boldsymbol{r^A_{B,C} r^A_{B,MB} + \delta}$	$\boldsymbol{r^A_{B,C}}$	**1**	
P	$\boldsymbol{\beta^A_{B,P} r^A_{B,MB} + \beta^A_{C,P}\,(r^A_{B,C} r^A_{B,MB} + \delta)}$	$\boldsymbol{\beta^A_{B,P} + \beta^A_{C,P}\, r^A_{B,C}}$	$\boldsymbol{\beta^A_{C,P} + \beta^A_{B,P}\, r^A_{B,C}}$	**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. The factor δ is the extent to which the C to MB correlation differs from the spurious correlation due to the correlation between C and B and between B and MB.

The factor δ is the extent to which the C to MB correlation differs from the spurious correlation due to the correlation between C and B and between B and MB. 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. Fitting any of the Models shown here is equivalent to obtaining the best fit between the implied and sample covariance and correlation matrices.

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.***

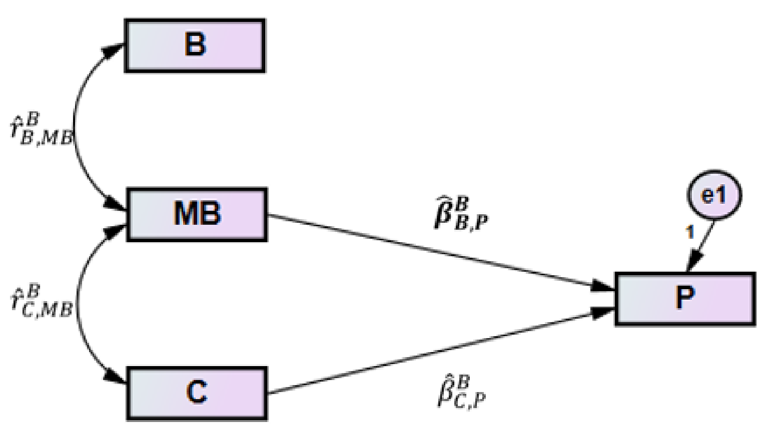

Figure A3. Model B: This is effectively Model C (or equivalently Model A) in which the positions of B and MB have been interchanged. Aside from the inclusion of B, this is also a very close representation of the Ordinary Least Squares equation used by Robinson (2013) for certain GLMs where P is the target 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. <u>Inclusion of this covariance term leads to results which do not differ from the results obtained from Model OLS in Fig. 7.</u> 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 $\hat{\beta}^B_{B,P}$ should strictly be denoted $\hat{\beta}^B_{MB,P}$, however the former representation is retained since it is regarded as the best estimate of the latter.

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

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

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 this 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	$r^A_{B,C}r^A_{B,MB}+\delta$	$r^A_{B,C}$	1	
P	$\hat{\beta}^B_{B,P}+\hat{\beta}^B_{C,P}\,(r^A_{B,C}r^A_{B,MB}+\delta)$	$\hat{\beta}^B_{B,P}r^A_{B,MB}+\hat{\beta}^A_{C,P}\,r^A_{B,C}$	$\hat{\beta}^B_{B,P}\left(r^A_{B,C}r^A_{B,MB}+\delta\right)+\hat{\beta}^B_{C,P}$	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:

$$\hat{\beta}^B_{B,P}+\hat{\beta}^B_{C,P}\left(r^A_{B,C}r^A_{B,MB}+\delta\right)=\beta^A_{B,P}r^A_{B,MB}+\beta^A_{C,P}\,(r^A_{B,C}r^A_{B,MB}+\delta)$$

$$\hat{\beta}^B_{B,P}\left(r^A_{B,C}r^A_{B,MB}+\delta\right)+\hat{\beta}^B_{C,P}=\beta^A_{C,P}+\beta^A_{B,P}r^A_{B,C}$$

These equations related 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 $\beta^A_{B,P}$, δ, and the true impact of pelagic catch on penguins $\beta^A_{C,P}$. The solution is as follows:

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

$$\hat{\beta}^B_{C,P}=\frac{\beta^A_{C,P}+\beta^A_{B,P}r^A_{B,C}-\left(r^A_{B,C}r^A_{B,MB}+\delta\right)(\beta^A_{B,P}r^A_{B,MB}+\beta^A_{C,P}(r^A_{B,C}r^A_{B,MB}+\delta))}{1-(r^A_{B,C}r^A_{B,MB}+\delta)^2}$$

The behavior of the estimators $\hat{\beta}^B_{B,P}$ and $\hat{\beta}^B_{C,P}$ can be explored for selected values of $r^A_{B,C}$, $r^A_{B,MB}$, $\beta^A_{B,P}$ **, δ,** and $\beta^A_{C,P}$.

The reliability of these two formulae was first tested using SEM software (Analysis of Moment Structures (AMOS) from IBM SPSS) to (a) fit the Fig. A1 model to the correlation matrix to verify that the true values are obtained, and then (b) to fit the Fig. A3 model to the same correlation matrix to verify that the standardized regression weights match those provided by the formulae above. Since there is an element of potential circular hidden error *in this process, a simulation exercise was carried out. 100 000 realizations of standardized data were generated from a multivariate normal distribution with a covariance matrix equal to the correlation matrix used for the SEM software based test (the R routine mvrnorm() was used) and fits were then carried out using a standard multiple linear regression package. Agreement was achieved between all three methods (SEM software, algebraic result, R simulations and linear regression).*

Selected results are reported below in Tables A1 –A4 and Figures A6 – A8:

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

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

$\boldsymbol{r}^{A}_{B,MB}$	$\boldsymbol{r}^{A}_{B,C}$	$\boldsymbol{\beta}^{A}_{B,P}$	$\boldsymbol{\beta}^{A}_{C,P}$	$\hat{\boldsymbol{\beta}}^{B}_{C,P}$	$\hat{\boldsymbol{\beta}}^{B}_{B,P}$
0.400	0.300	0.800	0.28	0.484	0.295
0.600	0.300	0.800	0.28	0.438	0.451
0.800	0.300	0.800	0.28	0.372	0.618
0.800	**0.500**	0.800	0.28	0.451	0.571
0.600	**0.200**	0.800	0.28	0.383	0.468

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. The value δ = 0 was used.

$r_{B,MB}^{A}$	$r_{B,C}^{A}$	$\beta_{B,P}^{A}$	$\beta_{C,P}^{A}$	$\widehat{\beta}_{C,P}^{B}$	$\widehat{\beta}_{B,P}^{B}$
0.400	0.300	0.800	**-0.28**	-0.080	0.295
0.600	0.300	0.800	**-0.28**	-0.121	0.451
0.800	0.300	0.800	**-0.28**	-0.188	0.618
0.800	**0.500**	0.800	**-0.28**	-0.109	0.571
0.600	**0.200**	0.800	**-0.28**	-0.176	0.468

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. The value δ = 0 was used.

$r_{B,MB}^{A}$	$r_{B,C}^{A}$	$\beta_{B,P}^{A}$	$\beta_{C,P}^{A}$	$\widehat{\beta}_{C,P}^{B}$	$\widehat{\beta}_{B,P}^{B}$
0.400	-0.300	0.800	0.28	-0.484545	0.295455
0.600	-0.300	0.800	0.28	-0.438743	0.451426
0.800	-0.300	0.800	0.28	-0.371681	0.617997
0.800	**-0.500**	0.800	0.28	-0.451429	0.571429
0.600	**-0.200**	0.800	0.28	-0.383896	0.467532

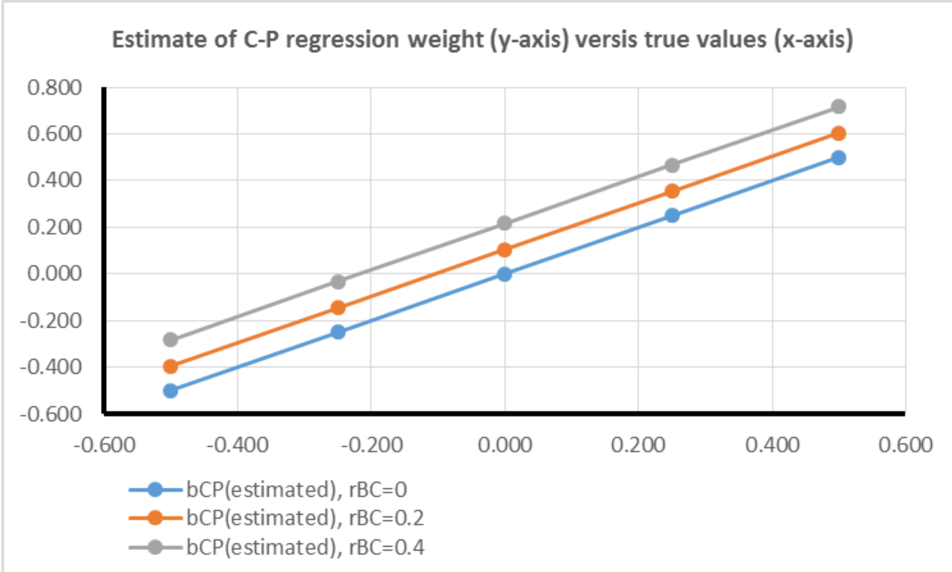

Figure A6. Estimates of $\widehat{\beta}^{B}_{C,P}$ (y-axis) for ranges of values of $r^{A}_{B,C}$ (different coloured lines), as a function of the true value $\beta^{A}_{C,P}$ (x-axis) for $r^{A}_{B,MB}$ = 0.6 and for $\beta^{A}_{B,P}$= 0.8. For these results δ = 0.00.

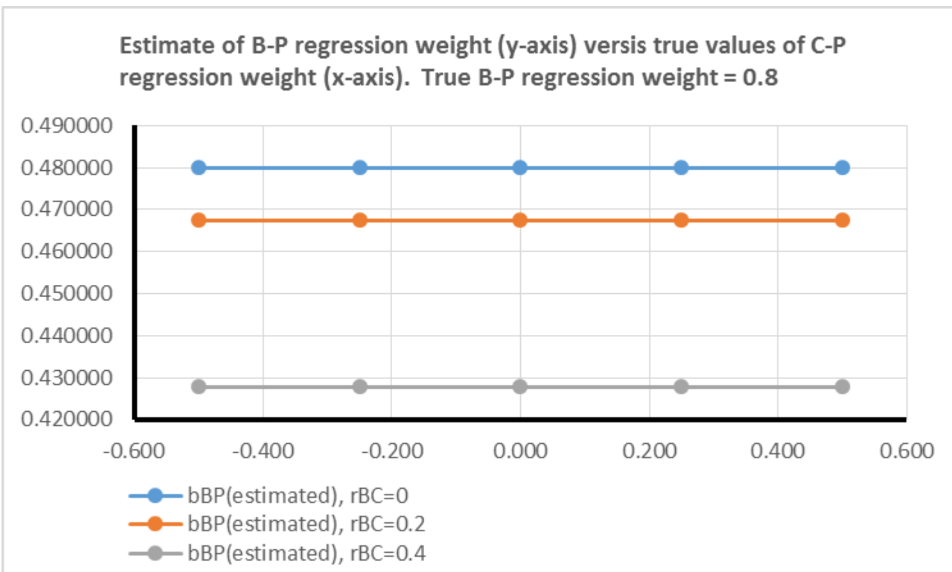

Figure A7. Estimates of $\widehat{\beta}^{B}_{B,P}$ (y-axis) for ranges of values of $r^{A}_{B,C}$ (different coloured lines), as a function of the true value $\beta^{A}_{C,P}$ (x-axis) for $r^{A}_{B,MB}$ = 0.6 and for $\beta^{A}_{B,P}$= 0.8. For these results δ = 0.00.

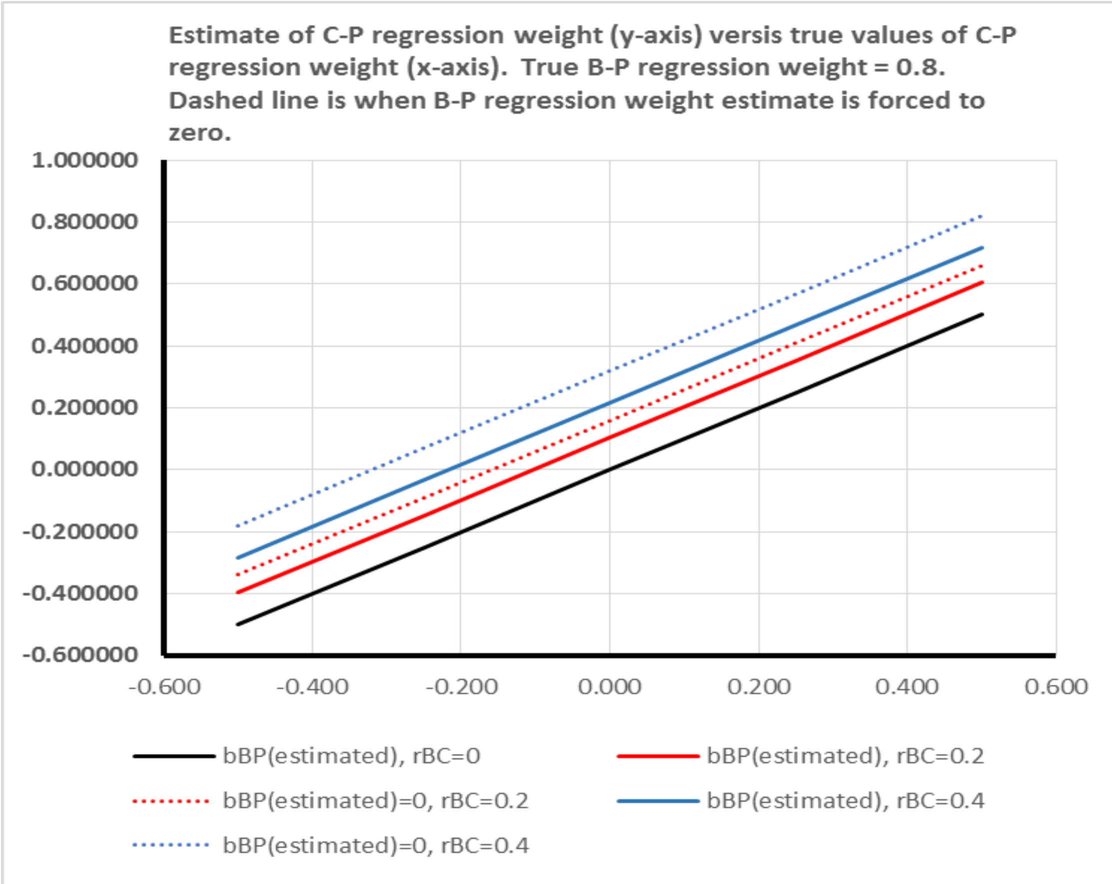

Figure A8. Estimates of $\widehat{\beta}^{B}_{C,P}$ (y-axis) for different values of $r^{A}_{B,C}$ (different lines), as a function of the true value $\beta^{A}_{C,P}$ (x-axis) for $r^{A}_{B,MB}$ = 0.6 and for $\beta^{A}_{B,P}$= 0.8, and when $\widehat{\beta}^{B}_{B,P}$ is either estimated (lower of 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. For these results δ = 0.00

<u>WRIGHT'S LAWS</u>

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.