Some investigations of the regression approach of FISHERIES/2020/JUL/SWG-PEL/60 to use the 2020 recruitment survey result to provide an estimate for the strength of the most recent west coast sardine recruitment

D.S. Butterworth and A. Ross-Gillespie[1]

Summary

Further investigation of statistical diagnostics for the recruit survey *vs* November model estimates regression for west coast sardine suggest that the estimate of the latter for the most recent recruitment be increased somewhat from 17.47 to 21.09, but that the estimate of the associated standard deviation in log space be maintained at 0.635.

Introduction

This document presents results for some further analyses of the regression of the historical survey results for year *y* (*surv(y)*) against the historical assessment estimates of November recruitment (*R(y)*) in calendar year *y*-1. Butterworth and Ross-Gillespie (2020) presented results for regressions using the data for 1985-2019 as well as for 2005-2019. The reason for using the reduced data set was concern over trends in the residuals for the regression applied to the entire data set. Subsequent to the presentation of Butterworth and Ross-Gillespie (2020) to the Working Group, however, a glitch in the excel cell-referencing for the plot in Figure 2(b) (this plotted the residuals of the fit of Equation (1) below for the data spanning the whole period of 1985-2019) was found;. Figure 1 of this document shows the corrected plot of the residuals, which no longer indicates the previous (but incorrect) systematic trend in these residuals, suggesting that applying the regression to the whole 1985-2019 period could be acceptable. Note that the glitch affected that Figure 2(b) plot only, and not any of the other results presented in Butterworth and Ross-Gillespie (2020). A corrected version of Butterworth and Ross-Gillespie (2020) will be circulated in due course.

That acceptability does, however, still need to be formally checked. Furthermore, there is potential concern about the appropriateness of the estimate of variance[2] when the regression is applied to the entire data set, as there seem to be indications that variability about the relationship regressed is higher in later years, and those years are more pertinent to any inference about the variance of any estimate of the most recent recruitment *R*(2020). In order to investigate this further, the regression of Equation (1) below is applied to a range of sub-sets of the 1985-2019 data set, starting with the whole set and then incrementally increasing the starting year, i.e. the regression is applied to the 1986-2019 data, then the 1987-2019 data and so on, to the set corresponding to the 2010-2019 data. Key estimates for the regression for each starting year are reported in this document.

As for Butterworth and Ross-Gillespie (2020), the relationship between *surv(y)* and *R(y)* is assumed to be:

$$ln\ surv(y) = ln\ R(y) + ln\ k + eps(y) \qquad eps(y) \sim \mathrm{N}(0, sig^2) \qquad (1)$$

Results and Discussion

Figure 1 plots the residuals when Equation (1) is applied to the entire data set. Figure 2 shows the estimated *k* values from Equation (1) when this Equation is applied to the data with a range of different starting years. Figure 3 plots the point estimates and 90% CIs for *sig* of Equation (1), also for this range of different starting years. Figure 4 plots the same *sig* values and 90% CIs, but for the case when the data set is split into two: Equation (1) is first applied to the earlier data from 1985-2001 and then to the second half of the data set from 2002-2019. Figure 5 plots the slope of the residuals for each analysis for the range of different starting years, and Figure 6 plots the value of November recruitment

[1] Marine Resource Assessment and Management Group, Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch, 7701.

[2] This estimate of variance is used to update prior distributions in the calculations of Butterworth and Ross-Gillespie (2020).

(estimated from Equation (1)) that corresponds to the 2020 recruitment survey result of 7.01 for each starting year assumed.

The CIs for slope shown in Figure 5 indicate that none of these values is significantly different from zero, suggesting that there is no problem to use the whole time series to estimate *k* and thence *R*(2020). This would replace the estimate of 17.47 in Butterworth and Ross-Gillespie (2020) by 21.09.

Use of the corresponding value of *sig* of 0.499 for this whole period does, however, seem problematic as there are clear indications that *sig* may be smaller earlier and larger later in the period, as shown in both Figures 3 and 4. Accordingly it seems better to retain the existing estimate for *sig* of 0.635 which was based on the period commencing in 2005, both for the *R*(2020) estimate itself, and for updating priors for this estimate. This choice of 2005 is arbitrary to some extent, but does seem a reasonable compromise between making allowance for this variance being larger later in the series, but being estimated with increasingly poorer precision as successively fewer years are taken into account in such estimation.

Reference

Butterworth, D.S. and Ross-Gillespie, A. 2020. Continuous approach applied to sardine recruitment survey estimate to advise on juvenile sardine TAB estimates for 2020 and the corresponding recruitment strength. FISHERIES branch document FISHERIES/2020/JUL/SWG-PEL/60.

Table 1: Output statistics for regression of the survey recruitment estimate (*surv*) against model-estimated recruitment (*R*) using Equation (1). The regression is conducted for a range of starting years, i.e. the regression for the 1985 row below uses the data from 1985-2019, the 1989 row the data from 1989-2019 and so on, with year referring to the year in which the recruitment survey took place, so that the model estimate refers to November of the previous year. Estimates for *k* and *sig* from Equation (1) are listed, as is the slope of the residuals (resid=ln(surv)-ln(surv_hat)) against year and the standard error of this estimate of the slope. Expected *R* gives the value of *R* that would be expected for the 2020 survey estimate of 7.01 when solving for *R(y)* using Equation (1) for the estimated *k* value.

			Slope of residuals		
Starting year	*k*	*sig*	Estimate	(se)	Expected *R*
1985	0.332	0.499	0.014	(0.008)	21.09
1986	0.333	0.507	0.015	(0.009)	21.05
1987	0.333	0.515	0.016	(0.010)	21.05
1988	0.332	0.523	0.019	(0.010)	21.12
1989	0.338	0.524	0.017	(0.011)	20.77
1990	0.348	0.506	0.012	(0.011)	20.15
1991	0.349	0.515	0.013	(0.012)	20.11
1992	0.351	0.524	0.014	(0.013)	19.99
1993	0.356	0.528	0.012	(0.014)	19.70
1994	0.357	0.539	0.012	(0.015)	19.64
1995	0.357	0.550	0.014	(0.016)	19.62
1996	0.353	0.560	0.019	(0.017)	19.85
1997	0.357	0.571	0.018	(0.019)	19.66
1998	0.349	0.576	0.027	(0.020)	20.08
1999	0.360	0.572	0.022	(0.022)	19.46
2000	0.373	0.565	0.014	(0.024)	18.78
2001	0.375	0.581	0.015	(0.026)	18.71
2002	0.374	0.598	0.019	(0.030)	18.77
2003	0.377	0.617	0.020	(0.033)	18.62
2004	0.386	0.630	0.013	(0.038)	18.15
2005	0.401	0.635	-0.001	(0.043)	17.47
2006	0.420	0.638	-0.022	(0.047)	16.70
2007	0.409	0.659	-0.015	(0.055)	17.12
2008	0.418	0.688	-0.030	(0.065)	16.79
2009	0.445	0.691	-0.077	(0.071)	15.76
2010	0.435	0.729	-0.089	(0.088)	16.13

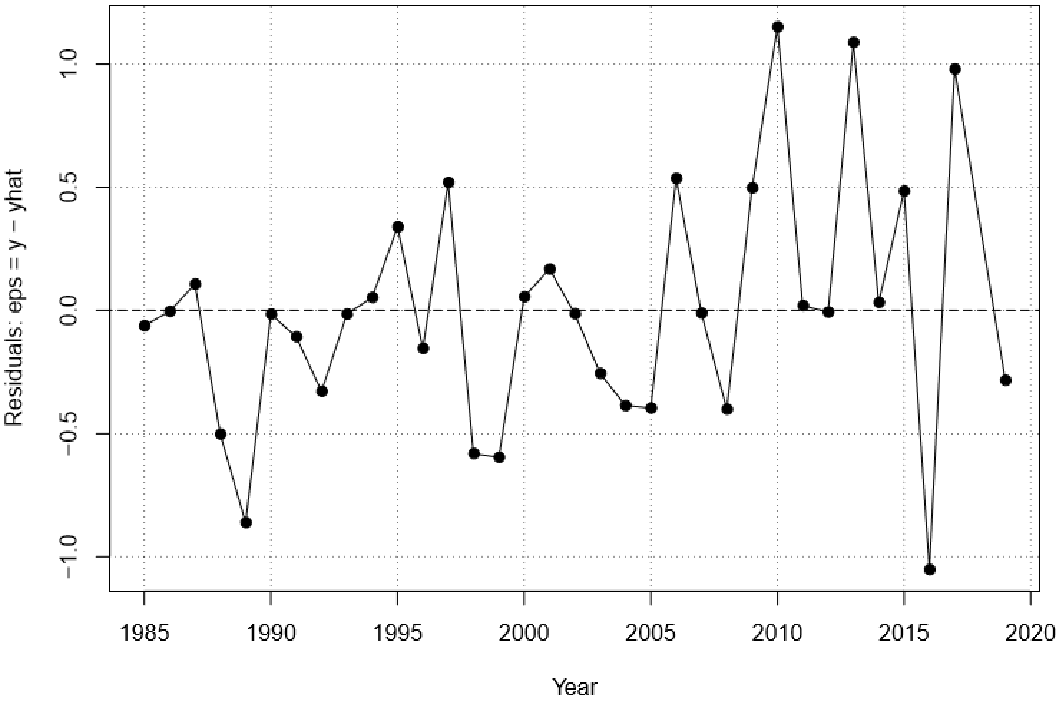

Figure 1: Plots of the residuals from Equation (1) when the data spanning the entire period of 1985-2019 are used. The residuals are defined as *y-y_hat*, where *y=surv(y)*.

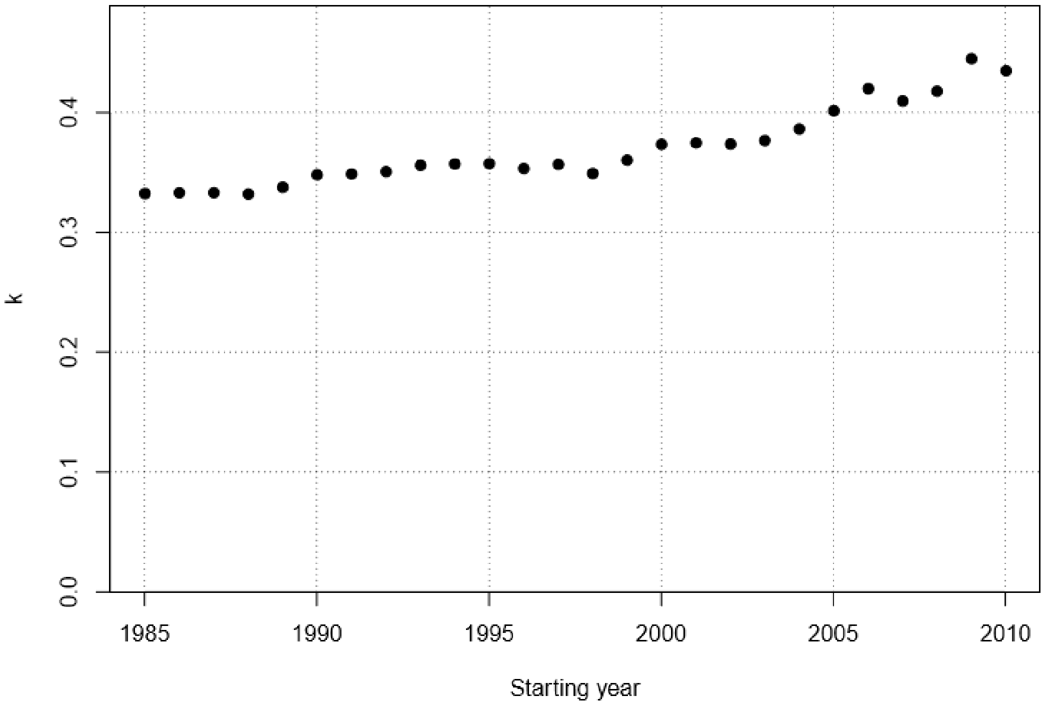

Figure 2: Estimated *k* values from Equation (1) for a series of regressions where the starting year for the data used is increased incrementally. The *k* estimate for 1985 is, therefore, the estimate that arises when the data set from 1985-2019 is used, the 1986 *k* estimate is for the data set spanning 1986-2019 and so on.

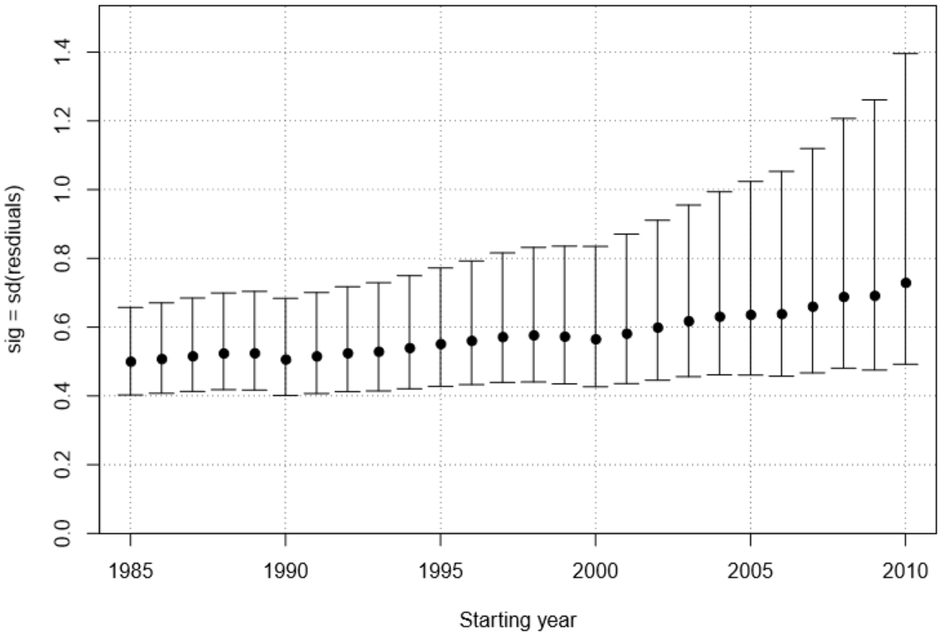

Figure 3: Point estimates and 90% confidence intervals (CI) for *sig* from Equation (1) are shown for the same range of starting years as described in Figure 2. The value of *sig* is the standard deviation of the residuals, i.e. $sig = \sqrt{\sum(x_i - \mu)^2/(n-1)}$, where x_i is the residual for year *i.* The 90% CIs are calculated from a χ^2 distribution:

Lower limit: $\sqrt{(n-1)sig^2/(\chi^2_{\alpha/2})}$

Upper limit: $\sqrt{(n-1)sig^2/(\chi^2_{1-\alpha/2})}$

where *n* is the number of data points in the time series in question (i.e. the later the starting year, the fewer data points there will be) and the χ^2 distribution provides values for α of 0.05 to give a 90% CI for $n-1$ degrees of freedom.

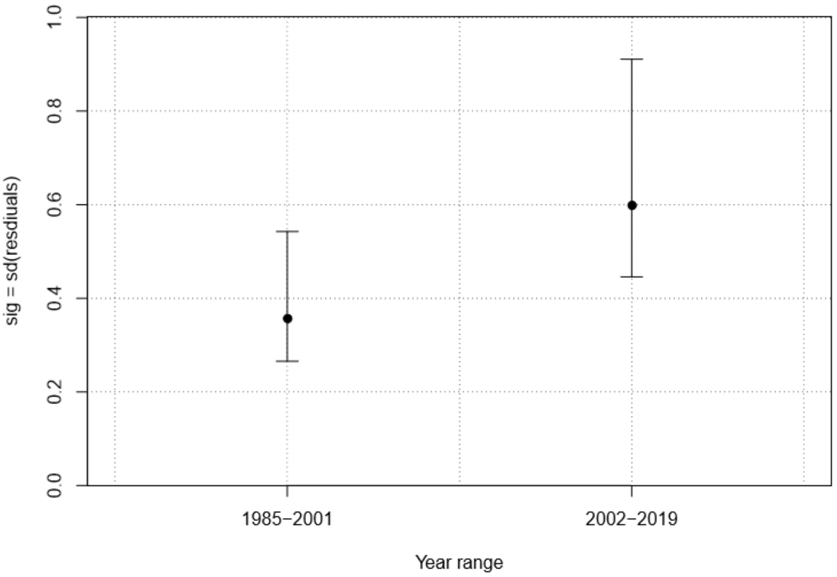

Figure 4: Estimates and 90% CIs of *sig* when the data set is split into two. The left point shows results for the regression of Equation (1) applied to the data from 1985-2001, and the right point when the data from 2002-2019 are used. The CIs are calculated in the same manner as described in the Figure 3 caption.

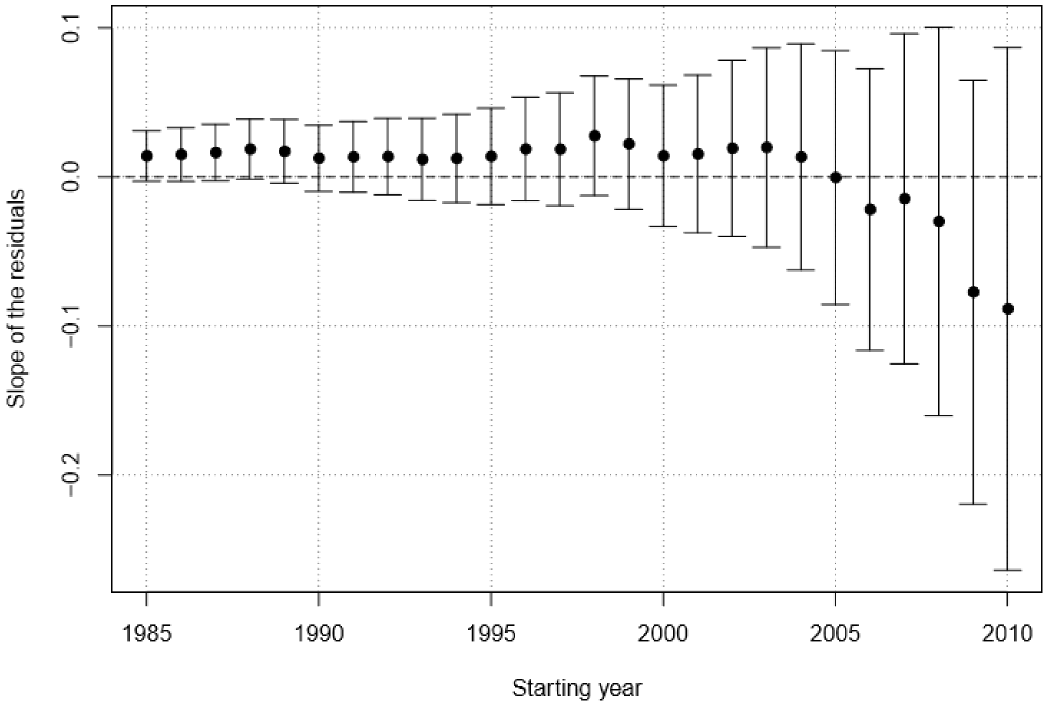

Figure 5: The residuals for each starting year are regressed against year to obtain an estimate and standard error of the slope of the residuals against year. The value plotted for 1985, for example, shows the slope and the approximate 95% confidence interval ($\pm 2se$) of the slope for the residuals when Equation (1) is applied to the entire data set over 1985-2019 (i.e. the slope of the regression line through the points in Figure 1). The value for 2010 on the other hand shows the slope and CI of the residuals when Equation (1) is applied to the data set over 2010-2019 only.

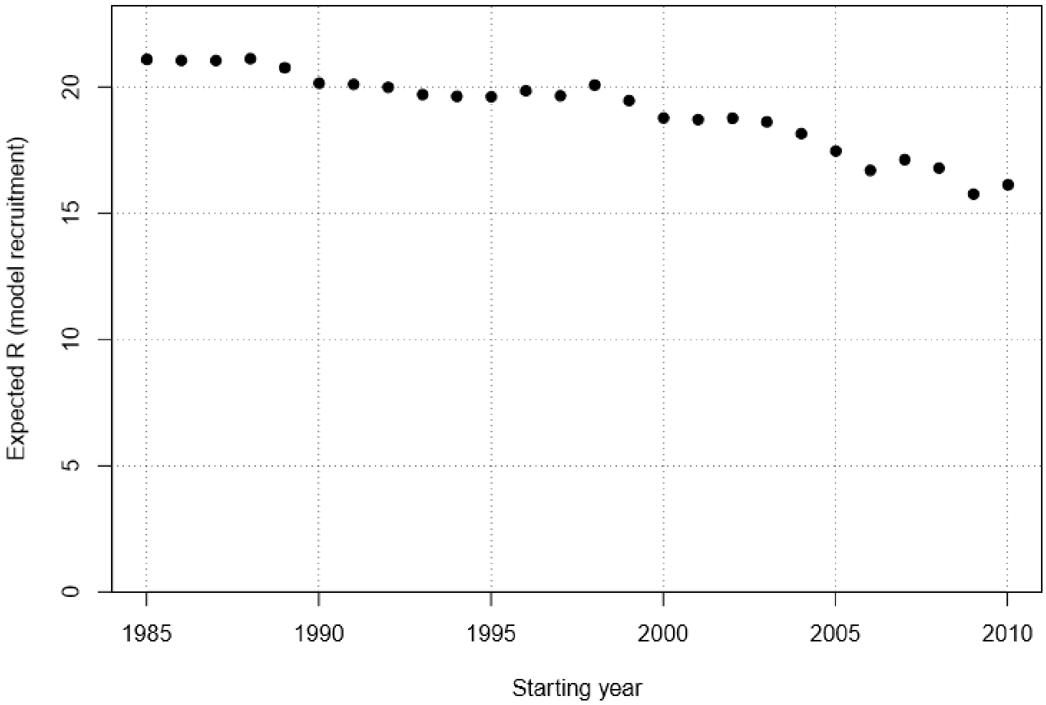

Figure 6: The expected assessment estimate of November recruitment given the 2020 survey result of 7.01 is shown for each starting year. These values are calculated by solving for *R(*2020*)* in Equation (1), using the *k* estimates from Figure 2.