A simulation study of impacts of error structure on modeling stock–recruitment data using generalized linear models

Yan Jiao, Yong Chen, David Schneider, and Joe Wroblewski

Abstract: Stock–recruitment (*S–R*) models are commonly fitted to *S–R* data with a least-squares method. Errors in modeling are usually assumed to be normal or lognormal, regardless of whether such an assumption is realistic. A Monte Carlo simulation approach was used to evaluate the impact of the assumption of error structure on *S–R* modeling. The generalized linear model, which can readily deal with different error structures, was used in estimating parameters. This study suggests that the quality of *S–R* parameter estimation, measured by estimation errors, can be influenced by the realism of error structure assumed in an estimation, the number of *S–R* data points, and the number of outliers in modeling. A small number of *S–R* data points and the presence of outliers in *S–R* data could increase the difficulty in identifying an appropriate error structure in modeling, which might lead to large biases in the *S–R* parameter estimation. This study shows that generalized linear model methods can help identify an appropriate error distribution in *S–R* modeling, leading to an improved estimation of parameters even when there are outliers and the number of *S–R* data points is small. We recommend the generalized linear model be used for quantifying stock–recruitment relationships.

Résumé : On a l'habitude d'ajuster les modèles stock–recrutement (*S–R*) aux données *S–R* à l'aide d'une méthode des moindres carrés. On présuppose que les erreurs de modélisation suivent une distribution normale ou lognormale, sans tenir compte si cette présupposition est réaliste ou non. Une simulation de type Monte Carlo nous a permis d'évaluer l'impact de la présupposition d'une structure d'erreur sur la modélisation *S–R*. Le modèle linéaire généralisé, qui peut s'accommoder facilement de diverses structures d'erreur, a servi à estimer les paramètres. Notre étude laisse croire que la qualité de l'estimation des paramètres *S–R*, reflétée dans les erreurs d'estimation, peut être influencée par le réalisme de la structure d'erreur choisie pour l'estimation, par le nombre de données de *S–R* et le nombre de données aberrantes dans la modélisation. Un petit nombre de données *S–R* et la présence de données aberrantes dans les données *S–R* peuvent rendre plus difficile l'identification d'une structure d'erreur appropriée dans la modélisation, ce qui peut mener à de fortes distorsions dans l'estimation des paramètres *S–R*. Notre étude démontre que les méthodes reliées au modèle linéaire généralisé permettent de définir une distribution d'erreur appropriée dans la modélisation *S–R*, ce qui mène à une meilleure estimation des paramètres, même lorsqu'il y a des données aberrantes et qu'il y a peu de données *S–R*. Nous recommandons l'utilisation du modèle linéaire généralisé pour quantifier les relations stock–recrutement.

[Traduit par la Rédaction]

Introduction

Stock–recruitment (*S–R*) models are mathematical functions that describe relationships between spawning stock abundance and subsequent recruitment. Future recruitment, which is essential for fisheries management, can be estimated from current spawners using *S–R* models (Hilborn and Walters 1992).

Large variations in recruitment have been observed for many fish species (e.g., Myers and Barrowman 1996; Power 1996; Hinrichsen 2001). The variations tend to increase with spawning stock biomass (Myers et al. 1995), which has led to wide adoption of the assumption that recruitment at a given level of spawning biomass follows a lognormal distribution (Hilborn and Walters 1992). This has been used as an alternative to the normal distribution (Hennemuth et al. 1980; Quinn and Deriso 1999; Hinrichsen 2001). The normal error distribution assumption is no longer widely used in *S–R* analyses, although it tends to be more realistic if the survival of individuals during their early life stages is density-independent and constant (Shelton 1992).

Error structure is an integral part of modeling (Carroll and Ruppert 1984; Schnute 1991; Chen and Paloheimo 1998). The impact of unrealistic assumptions of model error struc-

Received 6 January 2003. Accepted 4 November 2003. Published on the NRC Research Press Web site at http://cjfas.nrc.ca on 3 March 2004.
J17266

Y. Jiao.[1] Biology Department, Memorial University of Newfoundland, St. John's, NF A1B 3X9, Canada.
Y. Chen. School of Marine Science, University of Maine, Orono, ME 04469, U.S.A.
D. Schneider and J. Wroblewski. Biology Department and Ocean Science Center, Memorial University of Newfoundland, St. John's, NF A1B 3X9, Canada.

[1] Present address: Department of Zoology, University of Guelph, Guelph, ON A1G 2W1, Canada (e-mail: yjiao@uoguelph.ca).

doi: 10.1139/F03-149

ture on parameter estimation has been evaluated with various fisheries models (Deriso et al. 1985; Bajdik and Schneider 1991; Cadigan and Myers 2001). Inappropriate error distribution assumptions can cause inaccurate estimates of model parameters, their variability, and the attained significance level of the fitted model. In this study, using a Monte Carlo simulation approach, we evaluated the importance of having a proper error distribution assumption and the effectiveness of using a generalized linear model in identifying a proper error distribution in modeling *S*–*R* data.

The generalized linear model (GzLM) is a maximum likelihood based method that provides a systematic framework by which parameters are estimated when the model error structure belongs to the exponential family. We considered several error structures in the exponential family, including normal, lognormal, gamma, and Poisson distributions. The likelihood of recruitment having normal or lognormal distributions was discussed in Shelton (1992) and Fogarty (1993). Lognormal and gamma distributions were used by Myers et al. (1995). The Poisson distribution is considered appropriate for counting data and for nonnegative and highly varied data (Bajdik and Schneider 1991; White and Bennetts 1996) and may be appropriate for recruitment.

Because the number of *S*–*R* data is often small (Hilborn and Walters 1992) and outliers are likely to be present in the *S*–*R* data (Chen and Paloheimo 1995; Hinrichsen 2001), the role of sample size and outliers in identifying appropriate error structure in *S*–*R* modeling was also evaluated in the simulation study.

The theory of GzLM was developed in the 1970s (Nelder and Wedderburn 1972) and later was expanded in theory and application during the 1980s and 1990s (McCullagh and Nelder 1983, 1989; Dobson 1990). The GzLM can be implemented by various software such as GLIM (Numerical Algorithms Group Ltd., Oxford, U.K.), SAS (SAS Institute Inc., Cary, N.C.), and SPLUS (MathSoft Engineering and Education Inc., Cambridge, Mass.). MATLAB (Math Works Inc., Natick, Mass.) was used in this study.

Materials and methods

Data simulation

We used the Cushing (eq. 1), Ricker (eq. 2), and Beverton–Holt (eq. 3) models (Quinn and Deriso 1999) as examples in modeling *S*–*R* data:

$$(1) \qquad R = \alpha S^{\beta}$$

$$(2) \qquad R = \alpha S e^{-\beta S}$$

$$(3) \qquad R = \alpha S/(1+\beta S)$$

where α and β are two parameters to be estimated, and S and R are the spawning stock biomass and its subsequent recruitment, respectively. The Ricker and Beverton–Holt models consider a density-dependent effect, but the Cushing model does not. In practice, the choice of these models depends on the data. In this study, we considered the Cushing model first. Two scenarios were used in simulating *S*–*R* data with eq. 1. One scenario was to randomly sample S data from 1000 to 10 000 values using a uniform distribution and then calculate corresponding R values according to eq. 1 and an assumed error structure, which includes normal, lognormal, gamma, and Poisson functions (Table 1). The values of 2.012 and 0.857 were used for $\ln(\alpha)$ and β, respectively, in simulating the R data from the S data, which were taken from the pink salmon (*Oncorhynchus gorbuscha*) fishery in northern southeast Alaska (Quinn and Deriso 1999). For the second scenario, the *S*–*R* data were simulated using actual S data of pink salmon in northern southeast Alaska using the four types of error distributions listed in Table 1 and the two parameters for the pink salmon described above. The data simulated in the first and second scenarios are referred to as the first and second data sets, respectively, in this paper.

We then repeated the above approach for the Ricker and Beverton–Holt models (i.e., eqs. 2 and 3). The values used for the two parameters were $\ln(\alpha) = 1.047$ and $\beta = 5.52 \times 10^{-5}$ in the Ricker model and $\ln(\alpha) = 1.042$ and $\beta = 5.92 \times 10^{-5}$ in the Beverton–Holt model, also taken from the pink salmon stock.

Generalized linear model (GzLM)

The generalized linear model is sometimes abbreviated as GLM (McCullagh and Nelder 1989; Lindsey 1997; Myers et al. 2001), GLIM (software distributed by Numerical Algorithms Group), or GLZ (StatSoft Inc., Tulsa, Okla.). In this paper, we use GzLM to differentiate the generalized linear model from the general linear model (GLM) and the software GLIM.

A GzLM has three components (McCullagh and Nelder 1989). One is the random component $\boldsymbol{Y}$, which is a vector of observations y having n components that are independently distributed with means $\boldsymbol{\mu}$. The second is the systematic component, which is a specification for the vector $\boldsymbol{\mu}$ in terms of a small number of unknown parameters $\beta_1, \beta_2, \ldots, \beta_p$. A linear predictor $\boldsymbol{\eta}$ is given by $\boldsymbol{\eta} = \sum_{j=1}^{p} \mathbf{X}_j \beta_j$, where $\mathbf{X}$ is the model matrix or the covariates for observation $\boldsymbol{Y}$. The third component is the link between the random and systematic components. It is often written as $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, where g is the link function (McCullagh and Nelder 1989). In the case of the GLM, $\boldsymbol{\eta} = \boldsymbol{\mu}$ (i.e., identity link). Thus GLM is a special case of GzLM. The likelihood function of $\boldsymbol{Y}$ can be written as

$$(4) \qquad L = \prod_{1}^{n} f(y;\theta)$$

where f is the probability distribution function (pdf) which depends on the parameter(s) θ. If the pdf is a member of the exponential class, when $\boldsymbol{Y}$ is a discrete type, then we could have

$$(5) \qquad \begin{aligned} f(y;\theta) &= \exp[p(\theta)K(y) + s(y) + q(\theta)], \quad y = a_1, a_2, a_3, \ldots \\ &= 0, \quad \text{elsewhere} \end{aligned}$$

The log-likelihood function is then

$$(6) \qquad \mathrm{LL} = p(\theta)\sum_{1}^{n} K(y_i) + \sum_{1}^{n} S(y_i) + nq(\theta)$$

and where $\sum_{1}^{n} K(y_i)$ is a sufficient statistic for the parameter θ and $p(\theta)$ is the canonical link for a distribution for which the pdf is $f(y;\theta)$. Here, $\boldsymbol{Y}$ is recruitment R, $\mathbf{X}$ is spawning

Table 1. The distributions, their corresponding canonical links, and links used in this paper.

Distribution	Canonical link	Links used in this paper: Cushing and Ricker models	Links used in this paper: Beverton–Holt model
Normal	$\boldsymbol{\eta} = \boldsymbol{\mu}$	$\boldsymbol{\eta} = \log(\boldsymbol{\mu})$	$\boldsymbol{\eta} = 1/\boldsymbol{\mu}$
Lognormal (log-transformed data)	$\boldsymbol{\eta} = \boldsymbol{\mu}$	$\boldsymbol{\eta} = \boldsymbol{\mu}$	
Gamma	$\boldsymbol{\eta} = 1/\boldsymbol{\mu}$	$\boldsymbol{\eta} = \log(\boldsymbol{\mu})$	$\boldsymbol{\eta} = 1/\boldsymbol{\mu}$
Poisson	$\boldsymbol{\eta} = \log(\boldsymbol{\mu})$	$\boldsymbol{\eta} = \log(\boldsymbol{\mu})$	$\boldsymbol{\eta} = 1/\boldsymbol{\mu}$

stock biomass S, and θ represents the parameters in the S–R models.

The exponential family includes the normal, Poisson, gamma, and other distributions. The link function $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ relates the mean of the response variable $\mathbf{Y}$ to the linear combination of the $\mathbf{X}_i$. Common choices of link functions include identity, logarithmic, reciprocal, power, and logit (McCullagh and Nelder 1989; The Math Works Inc. 2002). The GzLM is flexible to incorporate different links and not be limited to the canonical links. The links that we used in this paper were based on the S–R model structure, because we did not consider changing the S–R model form. The choice of the link functions did not affect the assumption about the distribution of $\mathbf{Y}$ in GzLM. The distributions, their corresponding canonical links, and links used in this paper are shown in Table 1. The GzLM model was used to estimate the parameters given the link function $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ according to the maximum log-likelihood method (i.e., eq. 6). The parameters were then used to calculate the expected value of the response variable (i.e., recruitment) and then the residuals. Homogeneity of residuals was evaluated and used as a criterion to determine if the model error structure was appropriate. When different model error assumptions were used in a GzLM, the one that resulted in homogeneous residuals was considered as the most appropriate one (McCullagh and Nelder 1989). An example resulting from the application of GzLM with the four error distributions to S–R data simulated with a normal distributed error is shown (Fig. 1). This example suggests that residuals tend to be homogeneous when error distribution is correctly defined in the GzLM.

The GzLM parameter estimations for stock–recruitment models

Cushing model

A log transformation in conjunction with a normal error assumption is commonly used in fitting the Cushing model to S–R data. This is equivalent to assuming that the untransformed recruitment has a lognormal distribution. With the lognormal transformation, the Cushing model (i.e., eq. 1) can be rewritten as

$$(7) \qquad \ln(R) = \ln(\alpha) + \beta\ln(S) + \varepsilon$$

where ε has a normal distribution defined as $N(0,\sigma^2)$; $\ln(R)$ is the dependent variable and $\ln(S)$ is the independent variable in the GzLM. The identity link $\boldsymbol{\eta} = \boldsymbol{\mu}$ is used, and the choice of error is normal. The parameter estimates are the same as those from the GLM model after the Cushing model is log transformed (i.e., eq. 7).

For error distributions that are normal, gamma, and Poisson, we can rewrite the Cushing model as

$$(8) \qquad E(R) = \alpha S^{\beta} = e^{\ln(\alpha)+\beta\ln(S)}$$

Model parameters are estimated based on the assumption of the distribution of R. In the GzLM analysis, R is the dependent variable and $\ln(S)$ is the independent variable. The log link is used, and the error choice is normal, gamma, and Poisson, respectively. The parameter estimates using a normal error distribution are the same as those estimated using a nonlinear least-squares method.

Ricker model

When the lognormal distribution is assumed for the error distribution of the Ricker model, eq. 2 can be rewritten as

$$(9) \qquad \ln(R) = \ln(\alpha) - \beta S + \ln(S) + \varepsilon$$

which is equivalent to

$$(10) \qquad R = S e^{\ln(\alpha)-\beta S} e^{\varepsilon}$$

where ε follows $N(0,\sigma^2)$; $\ln(R)$ is the dependent variable, negative S is the independent variable, and $\ln(S)$ is an offset in the GzLM. The identity link $\boldsymbol{\eta} = \boldsymbol{\mu}$ is used, and the choice of error is normal. The parameter estimates are the same as those from the GLM model after linearizing the Ricker model. The commonly used linearization method, which can be expressed as

$$(11) \qquad R/S = e^{\alpha-\beta S} + \varepsilon$$

is likely to introduce estimation errors because variable S appears in both sides of the equation (Quinn and Deriso 1999).

For error distributions that are normal, gamma, and Poisson, the Ricker model can be written as

$$(12) \qquad E(R) = e^{\ln(\alpha)-\beta S+\ln(S)}$$

Model parameters are estimated based on the assumption of the distribution of R. R is the dependent variable, negative S is the independent variable, and $\ln(S)$ is an offset in the GzLM. The log link $\boldsymbol{\eta} = \log(\boldsymbol{\mu})$ is used, and the choice of error is normal, gamma, and Poisson, respectively.

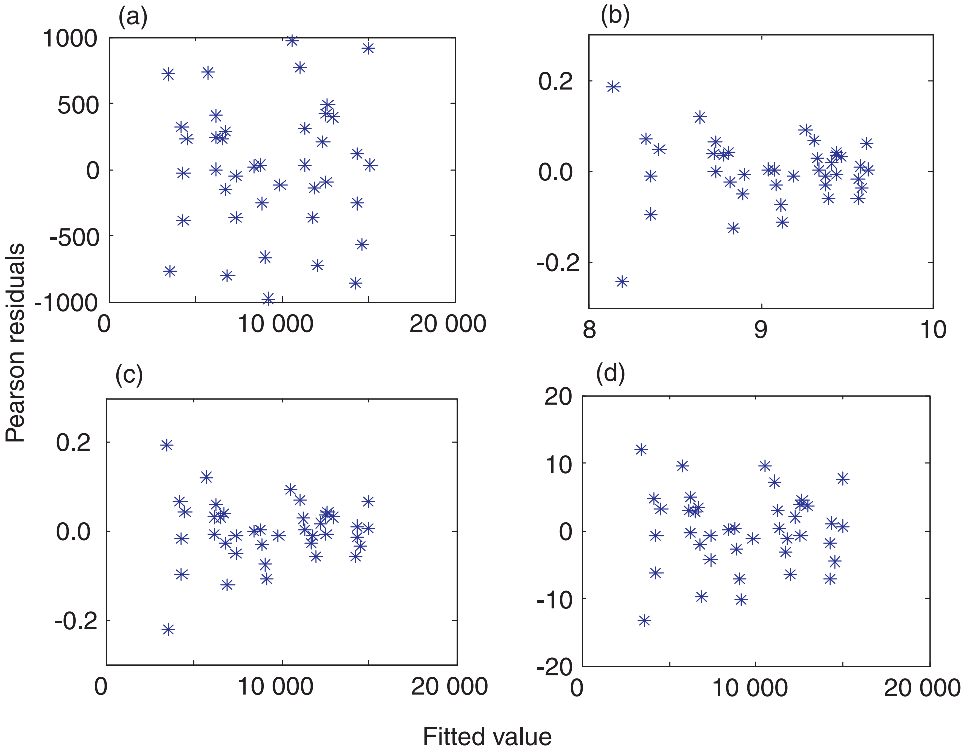
Beverton–Holt model

When the normal, gamma, or Poisson distribution is assumed for the error distribution of the Beverton–Holt model, eq. 3 can be rewritten as

$$(13) \qquad E(R) = \frac{1}{\dfrac{1}{\alpha S} + \dfrac{\beta}{\alpha}}$$

Model parameters are estimated based on the assumption of the distribution of R. R is the dependent variable and $1/S$ is the independent variable in the GzLM. The reciprocal link

Fig. 1. An example for showing the homogeneity and heterogeneity of residual distributions. Pearson residuals were used. The model error distribution used in simulating the stock–recruitment data is normal; the model error distributions used in the generalized linear model for the parameter estimation are (*a*) normal, (*b*) lognormal, (*c*) gamma, and (*d*) Poisson.



$\eta = 1/\mu$ is used, and the choice of error is normal, gamma, and Poisson, respectively.

For error distributions that are lognormal, the Beverton–Holt model can be written as

$$(14)\qquad \ln(R) = \ln\frac{\alpha S}{1+\beta S} + \varepsilon = \ln(\alpha) + \ln(S) - \ln(1+\beta S) + \varepsilon$$

where ε is the error term following $N(0,\sigma^2)$. Because of the nonlinear parameters in the above equation, the code and method used to get the parameter estimate are complicated and approximated when using GzLM. Equation 14 can be regarded as a nonlinear equation with a normally distributed error structure, where ln(*R*) is the response variable and *S* is the independent variable. Considering the fact that the parameter estimate using GzLM in eq. 14 equals that using the nonlinear least-squares estimate, a nonlinear least-squares method was used in solving this equation instead of GzLM.

Simulation design

The following procedure was used in the simulation: (*i*) simulate *S*–*R* data sets using the Cushing model for error scenarios (listed in Table 2) with the total number of observations being 10, 20, and 40; and (*ii*) simulate *S*–*R* data sets with outliers by adding atypical errors to 10%, 20%, and 40% of the data, with the number of observations being 40. Both of the *S*–*R* data sets discussed previously in the Data simulation section were used in simulating data with and without outliers. The same approach was then applied to the Ricker model. The combination of three models (Cushing, Ricker, and Beverton–Holt), two data sets (randomly drawn *S* data and true *S* data for the pink salmon), four error distributions used in simulating the *S*–*R* data (normal, lognormal, gamma, and Poisson), four error distributions assumed in the GzLM for parameter estimation (normal, lognormal, gamma, and Poisson), three sample sizes (10, 20, and 40 *S*–*R* observations), and three different levels of outlier contaminations (10%, 20%, and 40%) at the sample size of 40 resulted in 480 simulation scenarios in total evaluated in this study. When adding outliers in the simulation study, if 10% of outliers were added, 90% of *S*–*R* data with the supposed model error structure were first simulated and then another 10% of the outliers with another model error structure were simulated. Finally, the 90% common data and 10% outliers were added together before the generalized linear model was used to estimate the parameters. For each simulation scenario, 1000 simulations were run to derive the stable results.

Departure of the estimated parameters from the true values was measured by the relative estimation bias (REB; eq. 15), relative estimate error (REE; eq. 16), and root mean square error (RMSE; eq. 17):

$$(15)\qquad \text{REB}\,(\%) = \frac{\left(\sum_{1}^{N} \beta_i^*\right) \Big/ N - \beta}{\beta} \times 100$$

$$(16)\qquad \text{REE}\,(\%) = \frac{\sum_{1}^{N} \left|\beta_i^* - \beta\right|}{N\beta} \times 100$$

Table 2. Comparison of the regression method proposed in this paper with the commonly used visual checking method in identifying the residual homogeneity for the 100 simulations (in this study, the Cushing model was used).

Model error distribution		Simulations with homogeneous residuals (%)					
		Sample size = 10		Sample size = 20		Sample size = 40	
in simulating *S–R* data	in GzLM estimation	Regression method	Visual checking	Regression method	Visual checking	Regression method	Visual checking
Normal	Normal	87	88	93	88	92	87
	Lognormal	70	68	40	48	10	16
	Gamma	72	70	43	47	10	13
	Poisson	83	77	80	80	40	40
Lognormal	Normal	82	57	65	40	20	18
	Lognormal	97	75	85	75	82	87
	Gamma	88	72	82	77	85	93
	Poisson	87	63	75	58	57	50
Gamma	Normal	82	47	57	43	33	27
	Lognormal	80	72	87	77	70	70
	Gamma	83	78	87	83	70	70
	Poisson	83	60	78	62	73	58
Poisson	Normal	83	62	70	62	43	40
	Lognormal	87	60	72	52	53	58
	Gamma	85	63	75	47	53	58
	Poisson	88	67	80	77	93	87

Note: The results derived for different sample sizes (10, 20, and 40) are compared. GzLM, generalized linear model; *S–R*, stock–recruitment.

$$(17)\qquad \mathrm{RMSE} = \left(\frac{\sum_{1}^{N} (\beta_i^* - \beta)^2}{N} \right)^{1/2}$$

where β_i^* is the estimated parameter value in the *i*th simulation, N is the number of simulations, and β is the true parameter value. The smaller the REB, REE, and RMSE values, the better the estimation approach performs.

Quantification of residual's homogeneity

Pearson residuals of the model fitting were calculated for each simulation scenario. Pearson residuals are the differences between observed and predicted values, standardized (divide by the estimated standard deviation of the fitted value) to make their variance (theoretically) constant. If the error distribution assumed in the estimation was consistent with the error distribution used in simulating the *S–R* data, then a plot of resultant Pearson residuals should show constant variances. Thus, to determine if the residuals are homogeneous, one can look at the residual diagnostic plot. A visual inspection of residual diagnostic plots is commonly used in evaluating the residual homogeneity, although the approach may be subjective, particularly when sample sizes are small. In this study, because 1000 simulations were run for each scenario to derive stable results, direct observation was laborious. Thus, we used the following quantitative method to evaluate the homogeneity of the residuals. (*i*) Regress the Pearson residual value (r) and the model predicted value ($\hat{R}$), i.e.,

$$(18)\qquad r = a_1 + b_1\hat{R}$$

Estimate the p value for b_1 to determine if it was significantly different from zero; $\alpha = 10\%$ was used. (*ii*) Regress the absolute Pearson residual value (r_2) and the model predicted value ($\hat{R}$), i.e.,

$$(19)\qquad r_2 = a_2 + b_2\hat{R}$$

Estimate the p value for b_2 to determine if it was significantly different from zero; $\alpha = 10\%$ was used. (*iii*) If both of the estimated p values were larger than $\alpha = 10\%$, we regarded the residuals as homogeneous, otherwise we regarded the residuals as not homogeneous. Many of the residuals show a right or left triangular-shaped pattern symmetric about the x axis when an inappropriate model error structure was used; therefore, we double-checked this situation by using both residuals and absolute value of residuals in the regression. This proposed method, referred hereafter as the regression method, was compared with the commonly used method, which involves visual examination of residual homogeneity, in 100 simulations to determine if both methods derived consistent conclusions in evaluating the residual homogeneity.

Results

The efficiency of the regression method in identifying the residual homogeneity was influenced by sample size (Table 2). When the sample size was 10, the percentage of the simulations with homogeneous residuals was 81% on average. This increased slightly with an increase in sample size (averaged 84% and 89% for sample sizes of 20 and 40, respectively; Table 2). A comparison of the regression method with the commonly used visual-checking method for the residual homogeneity suggests that the proposed regression method effectively identified the simulations that had homogeneous residuals.

When the error distributions used in the GzLM analysis were the same as those used in simulating *S–R* data, percent-

Table 3. Summary of the simulations in which the model error distributions used in the generalized linear model analyses were the same as those used in simulating stock–recruitment (*S*–*R*) data.

Model error distribution in simulating *S*–*R* data	No. of observations	Simulations with homogeneous residuals (%)	REB for ln(α) (%)	REB for β (%)	RMSE for ln(α)	RMSE for β
Normal	10	71.9	0.15	0.04	0.4398	0.0518
	20	79.1	0.02	0.01	0.2989	0.0351
	40	80.4	0.02	0.05	0.2106	0.0248
Lognormal	10	74.3	6.90	1.93	2.9511	0.3585
	20	79.4	0.01	0.19	1.9812	0.2413
	40	80.8	2.21	0.60	1.3260	0.1618
Gamma	10	71.3	0.61	0.15	0.4516	0.0542
	20	73.5	0.73	0.21	0.2915	0.0351
	40	79.1	0.37	0.10	0.1860	0.0224
Poisson	10	73.6	0.026	0.008	0.0724	0.0086
	20	78.5	0.026	0.008	0.0491	0.0059
	40	80.6	0.008	0.002	0.0342	0.0041

Note: The Cushing model and the first set of data, which had *S* values randomly drawn from 1000 to 10 000 values, were used in the simulation. Relative estimate bias (REB) and root mean square error (RMSE) for parameters are shown.

ages of the simulations that had homogeneous residuals increased with an increase in sample size (Table 3). The REEs and RMSEs for α and β tended to decrease with increasing sample size (Fig. 2; Table 3). The REBs of α and β for the sample size of 40 tended to be larger than those for the sample size of 20, but still smaller than for the sample size of 10. Because the REB is the difference between the mean parameter estimate in total simulation runs and the true value divided by the true value, the variance among the simulation runs was hidden. The REB estimate is unsteady compared with the REE and RMSE estimates, which accumulate error in every simulation run. (We did not show the REE estimates in the tables; we only show it in Fig. 2 as an example.)

When the error distributions used in the GzLM analysis differed from those used in simulating *S*–*R* data, percentages of the simulations with homogeneous residuals became smaller with an increase in sample size (Table 4). This was different from the observations made when model error distributions used in the GzLM analysis were the same as those used in simulating *S*–*R* data. The REBs of α and β increased with an increase in sample sizes when normal errors were used in simulating *S*–*R* data. For other errors used in simulating data, they decreased as the sample size increased (Table 4). The RMSEs of α and β decreased with increasing sample size.

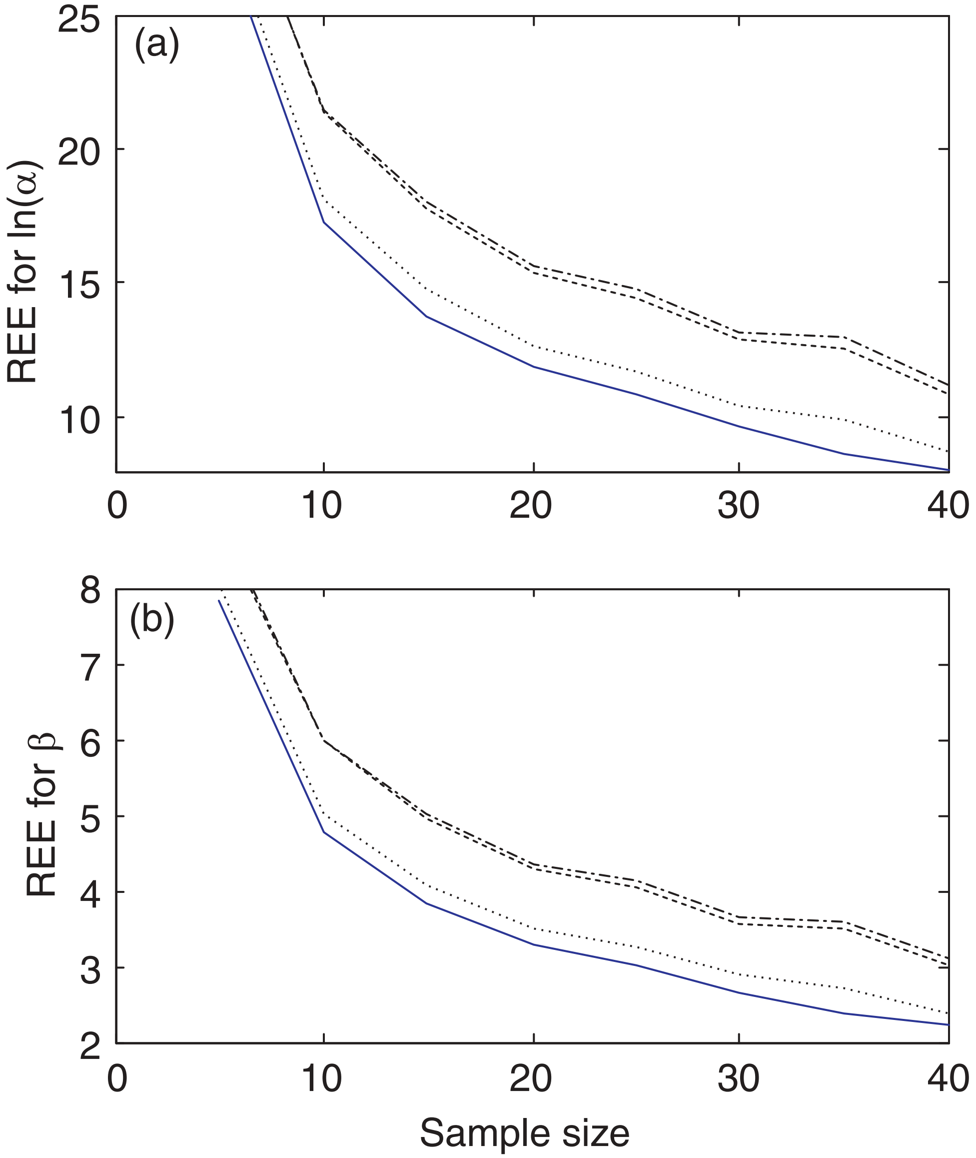
A comparison of Tables 3 and 4 indicates that percentages of the simulations with homogeneous residuals were always higher and REBs and RMSEs of α and β were always lower when the error assumptions used in the GzLM were the same as those used in simulating *S*–*R* data compared with those derived in the simulations when the model error distributions in the GzLM were not the same as the ones used in simulating data (Tables 3 and 4).

Different error assumptions in the GzLM yielded different estimates for the model parameters. Parameters estimated using lognormal and gamma distributions were similar, but parameters estimated using normal distribution in the GzLM were different from those estimated using lognormal and gamma distributions in the GzLM. Parameters estimated using a Poisson distribution tended to have values between those estimated using the normal error distribution versus lognormal and gamma distributions. Because this is a simulation study, we did not show the parameter estimate in every simulation. The trend of the mean error in total simulations can be observed from the REE and RMSE estimates (Fig. 2; Tables 3 and 4).

When the second *S*–*R* data (see Data simulation section) were used for the Cushing model, percentages of the simulations with homogeneous residuals were higher when the error distributions used in the GzLM were the same as those used in simulating *S*–*R* data (Table 5). Parameters estimated using the lognormal and gamma distributions were similar. Parameters estimated using the normal distribution in the GzLM differed from those estimated using the lognormal and gamma distributions. Parameters estimated using the Poisson distribution in the GzLM had values between those estimated using normal error and lognormal and gamma distributions in the GzLM. When the error distribution used in the GzLM was the same as that used in simulating *S*–*R* data, the percentages of the simulations with homogeneous residuals were higher and the REBs and RMSEs of α and β were smaller, and vice versa. This was consistent with the results derived for the first set of data, which had *S* values randomly drawn from 1000 to 10 000 values.

For the first set of data (randomly drawn *S* data) with outliers, when the error distributions used in the GzLM were the same as those used in simulating *S*–*R* data, the percentages of the simulations with homogeneous residuals decreased and the REBs and RMSEs of α and β increased with an increase in the number of outliers (Table 6). When the error distributions used in the GzLM were not the same as those used in simulating *S*–*R* data, the percentages of the simulations with homogeneous residuals increased with an increase in the number of outliers, if error distributions used in simulating *S*–*R* data were normal and error distributions used in the GzLM were lognormal and gamma. The same result could also be observed when the error distributions used in simulating *S*–*R* data were lognormal, gamma, and Poisson and the model errors used in the GzLM were normal. For other combinations of the distributions in simulating data and GzLM analyses, the percentages of simulations with ho-

Fig. 2. Relative estimation error (REE) for the two parameters (*a*) ln(α) and (*b*) β in the Cushing model in 1000 simulation runs using different sample sizes in the simulation study. The simulated model error distribution is normal; the model error distributions used in the parameter estimates are normal (solid line), lognormal (dashed–dotted line), gamma (dashed line), and Poisson (dotted line).



mogeneous residuals decreased with an increasing number of outliers (Table 7). The REBs and RMSEs of α and β increased with an increase in the number of outliers in most cases (Table 7). Similar conclusions could be obtained when outliers were present in the second set of data simulated based on pink salmon *S* data (Tables 8 and 9).

When the Ricker and Beverton–Holt models were applied to data simulated under different scenarios, the results derived were consistent with those described above for the Cushing model. For brevity, the results for those two models are not shown in this paper.

Discussion

The regression method proposed for checking the residual homogeneity was effective in identifying homogeneous distributions of residuals. However, the effectiveness decreased with sample size. This was consistent with the fact that visual observation for residual homogeneity was difficult when sample sizes were small. This suggests that the regression method that we used in checking residual homogeneity of a large number of simulations was effective. However, we would suggest using the visual observation method in *S–R* modeling when there is only one set of residuals output.

The simulation results show that GzLM can help identify an appropriate model error distribution for parameter estimation through a residual homogeneity analysis. In most cases of the simulation, the true model error distribution used in the GzLM provided the highest percentage of simulation runs with homogeneous residuals. According to the simulation, this (i.e., the use of an appropriate error distributions in the GzLM) can improve the parameter estimation.

Table 4. Summary of the simulations in which the model error distributions used in the generalized linear model analyses were not the same as those used in simulating stock–recruitment (S–*R*) data.

Model error distribution							
in simulating *S*–*R* data	in GzLM	No. of *S*–*R* observations	Simulations with homogeneous residuals (%)	REB for ln(α) (%)	REB for β (%)	RMSE for ln(α)	RMSE for β
Normal	Lognormal	20	26.3	2.57	0.70	0.4096	0.0486
		40	5.0	2.94	0.80	0.3033	0.0360
	Gamma	20	24.7	0.91	0.26	0.3999	0.0474
		40	4.6	0.84	0.23	0.2916	0.0346
	Poisson	20	55.5	0.31	0.09	0.3216	0.0378
		40	34.9	0.47	0.13	0.2315	0.0273
Lognormal	Normal	20	37.4	24.6	8.20	11.2407	1.2923
		40	13.1	5.25	3.02	2.2124	0.2666
	Gamma	20	81.8	4.39	0.45	2.0394	0.2480
		40	85.8	2.43	0.94	1.4105	0.1724
	Poisson	20	63.2	7.89	0.55	2.2546	0.2733
		40	43.9	2.72	0.85	1.5502	0.1890
Gamma	Normal	20	44.1	1.13	0.32	0.3392	0.0406
		40	17.8	0.20	0.05	0.2302	0.0277
	Lognormal	20	73.7	1.09	0.28	0.2924	0.0352
		40	64.2	0.09	0.01	0.1861	0.0224
	Poisson	20	72.6	0.85	0.24	0.2935	0.0353
		40	67.7	0.38	0.10	0.1915	0.0231
Poisson	Normal	20	58.6	0.041	0.012	0.0539	0.0064
		40	36.6	0.027	0.008	0.0367	0.0044
	Lognormal	20	58.5	0.046	0.013	0.0525	0.0063
		40	39.4	0.009	0.002	0.0371	0.0044
	Gamma	20	59.1	0.026	0.008	0.0525	0.0063
		40	37.9	0.016	0.005	0.0371	0.0044

Note: The Cushing model and the first set of data, which had *S* values randomly drawn from 1000 to 10 000 values, were used in the simulation. Relative estimate bias (REB) and root mean square error (RMSE) for parameters are shown. GzLM, generalized linear model; *S*–*R*, stock–recruitment.

Table 5. Summary of the simulations in which the Cushing model and the second set of data that had true *S* values of pink salmon were used in the simulation.

Model error distribution						
in simulating *S*–*R* data	in GzLM	Simulations with homogeneous residuals (%)	REB for ln(α) (%)	REB for β (%)	RMSE for ln(α)	RMSE for β
Normal	Normal	82.3	0.01	0.01	0.2144	0.0258
	Lognormal	26.9	2.18	0.59	0.3486	0.0425
	Gamma	31.1	0.37	0.10	0.3408	0.0415
	Poisson	59.6	0.01	0.01	0.2451	0.0296
Lognormal	Normal	37.1	7.86	0.80	2.8999	0.3556
	Lognormal	80.8	0.64	0.21	1.8501	0.2309
	Gamma	86.1	6.58	0.27	1.9416	0.2418
	Poisson	77.0	12.22	1.93	2.1647	0.2682
Gamma	Normal	32.8	0.43	0.12	0.3286	0.0405
	Lognormal	73.8	0.67	0.16	0.2644	0.0328
	Gamma	75.6	0.22	0.06	0.2636	0.0327
	Poisson	82.5	0.19	0.05	0.2639	0.0327
Poisson	Normal	53.0	0.011	0.003	0.0470	0.0058
	Lognormal	39.9	0.002	0.002	0.0472	0.0058
	Gamma	42.6	0.030	0.007	0.0472	0.0058
	Poisson	83.3	0.020	0.006	0.0417	0.0051

Note: Relative estimate bias (REB) and root mean square error (RMSE) for parameters were shown. GzLM, generalized linear model; *S*–*R*, stock–recruitment.

Table 6. Summary of the simulation in which the model error distributions used in the generalized linear model analyses were the same as those used in simulating stock–recruitment (*S*–*R*) data.

Model error distribution in simulating *S*–*R* data	Distribution that outliers follow	Data being outliers (%)	Simulations with homogeneous residuals (%)	REB for ln(α) (%)	REB for β (%)	RMSE for ln(α)	RMSE for β
Normal	Lognormal	10	66.8	1.77	0.67	0.7189	0.0866
		20	58.4	2.28	0.97	1.1859	0.1477
		40	43.4	5.83	2.29	1.5351	0.1852
Lognormal	Normal	10	80.0	4.08	1.17	1.2193	0.1485
		20	81.6	0.36	0.09	1.1835	0.1444
		40	79.5	0.52	0.16	0.9844	0.1203
Gamma	Normal	10	55.2	0.14	0.03	0.2030	0.0244
		20	47.9	1.37	0.38	0.2273	0.0273
		40	29.0	0.16	0.04	0.2332	0.0278
Poisson	Normal	10	68.4	0.12	0.03	0.0780	0.0092
		20	69.5	0.13	0.04	0.1044	0.0123
		40	59.5	0.14	0.04	0.1499	0.0177

Note: The Cushing model and the first set of data, which had *S* values randomly drawn from 1000 to 10 000 values, were used in the simulation. Outliers were added and the sample size was 40. Relative estimate bias (REB) and root mean square error (RMSE) for parameters are shown.

Table 7. Summary of the simulation in which the model error distributions used in the generalized linear model analyses were not the same as those used in simulating stock–recruitment (*S*–*R*) data.

Model error distribution							
in simulating *S*–*R* data	in GzLM	Data being outliers (%)	Simulations with homogeneous residuals (%)	REB for ln(α) (%)	REB for β (%)	RMSE for ln(α)	RMSE for β
Normal	Lognormal	10	56.6	1.21	0.31	0.4658	0.0563
		20	69.8	1.69	0.46	0.6479	0.0788
	Gamma	10	56.0	1.03	0.12	0.5111	0.0619
		20	68.8	0.15	0.31	0.7147	0.0870
	Poisson	10	70.4	0.25	0.10	0.5843	0.0664
		20	68.3	0.38	0.24	0.7996	0.0972
Lognormal	Normal	10	18.9	3.21	0.51	1.9861	0.2385
		20	22.0	1.38	1.66	1.9122	0.2305
	Gamma	10	83.2	7.83	0.72	1.2994	0.1572
		20	83.0	4.04	0.20	1.2773	0.1557
	Poisson	10	49.1	9.25	1.13	1.4226	0.1724
		20	54.6	4.76	0.01	1.4020	0.1704
Gamma	Normal	10	23.3	0.19	0.05	0.2329	0.0279
		20	30.6	1.20	0.33	0.2283	0.0273
	Lognormal	10	56.4	0.78	0.19	0.2042	0.0245
		20	47.0	2.17	0.58	0.2325	0.0278
	Poisson	10	75.5	0.17	0.04	0.2001	0.0241
		20	78.1	1.15	0.32	0.2100	0.0252
Poisson	Normal	10	64.5	0.16	0.04	0.0730	0.0086
		20	71.2	0.13	0.04	0.0952	0.0112
	Lognormal	10	50.3	0.33	0.09	0.1017	0.0121
		20	42.4	0.34	0.09	0.1347	0.0160
	Gamma	10	47.6	0.08	0.02	0.0983	0.0117
		20	42.4	0.11	0.03	0.1317	0.0157

Note: The Cushing model and the first set of data, which had *S* values randomly drawn from 1000 to 10 000 values, were used in the simulation. Outliers were added and the sample size was 40. Relative estimate bias (REB) and root mean square error (RMSE) for parameters are shown. GzLM, generalized linear model.

This study suggests that the number of *S*–*R* data can greatly influence the effectiveness of identifying correct error distributions in the GzLM analyses and estimation errors. A small number of *S*–*R* data is likely to lead to low effectiveness in identifying homogeneous residuals and large estimation errors in *S*–*R* modeling. Thus, we should be cautious in using the GzLM when analyzing a small number of *S*–*R* data.

The presence of outliers can also impact a GzLM analysis in identifying a correct error distribution. When the number of outliers was 10% of the data, the percentage of simulation

Table 8. Summary of the simulation in which the model error distributions used in the generalized linear model analyses were the same as those used in simulating stock–recruitment (*S*–*R*) data.

Model error distribution in simulating *S*–*R* data	Distribution that outliers follow	Data being outliers (%)	Simulations homogeneous residuals (%)	REB for ln(α) (%)	REB for β (%)	RMSE for ln(α)	RMSE for β
Normal	Lognormal	10	64.3	0.02	0.18	0.3168	0.0395
		20	40.2	18.44	11.32	3.1441	0.3871
		40	21.8	27.32	8.55	2.9112	0.3597
Lognormal	Normal	10	77.4	0.60	0.18	1.8214	0.2266
		20	67.3	4.57	1.29	1.5377	0.1895
		40	26.7	1.99	0.57	1.3729	0.1690
Gamma	Normal	10	76.2	0.22	0.06	0.2616	0.0323
		20	70.5	0.10	0.02	0.2411	0.0297
		40	61.0	0.63	0.18	0.2321	0.0286
Poisson	Normal	10	80.2	0.07	0.02	0.0590	0.0074
		20	69.1	0.08	0.02	0.1134	0.0142
		40	48.2	0.08	0.03	0.1209	0.0152

Note: The Cushing model and the second set of data that had true *S* values of pink salmon were used in the simulation. Outliers were added in simulating the data. Relative estimate bias (REB) and root mean square error (RMSE) for parameters are shown.

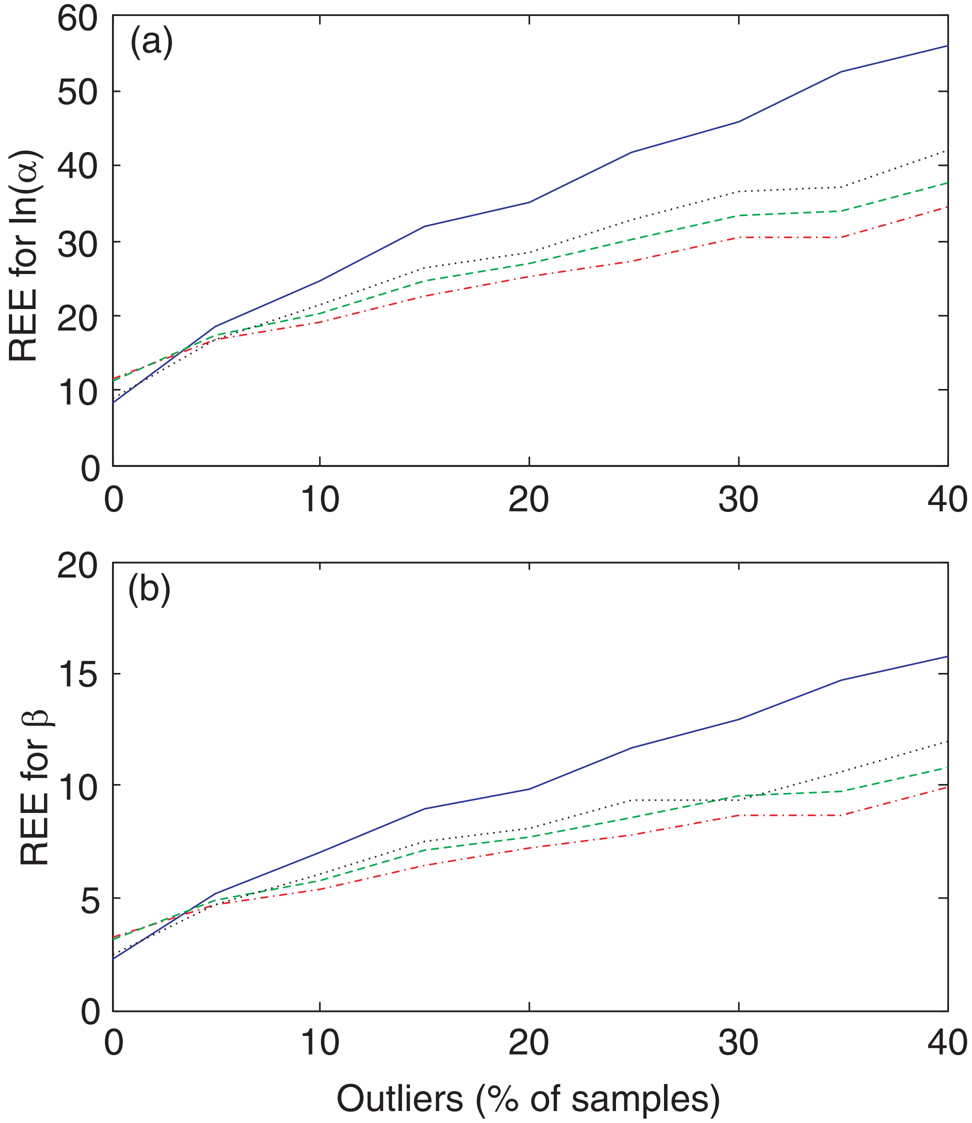
Table 9. Summary of the simulation in which the model error distributions used in the generalized linear model analyses were not the same as those used in simulating stock–recruitment (S–*R*) data.

Model error distribution							
in simulating S–*R* data	in GzLM	Data being outliers (%)	Simulations with homogeneous residuals (%)	REB for ln(α) (%)	REB for β (%)	RMSE for ln(α)	RMSE for β
Normal	Lognormal	10	83.2	1.71	0.43	0.5946	0.0760
		20	53.1	5.36	1.57	1.1047	0.1408
	Gamma	10	84.3	7.05	2.21	0.6534	0.0840
		20	53.5	18.04	5.66	1.2074	0.1549
	Poisson	10	73.8	4.05	1.33	0.4979	0.0641
		20	40.1	25.31	7.74	2.0522	0.2580
Lognormal	Normal	10	80.0	11.95	2.16	2.9611	0.3629
		20	86.1	41.40	10.64	1.3247	0.1527
	Gamma	10	82.1	14.52	2.72	1.9505	0.2412
		20	73.4	26.27	6.28	1.7398	0.2113
	Poisson	10	81.1	19.76	4.27	2.1775	0.2689
		20	85.7	35.73	9.03	1.4945	0.1772
Gamma	Normal	10	78.4	0.27	0.08	0.3213	0.0395
		20	83.0	0.07	0.03	0.1981	0.0241
	Lognormal	10	74.4	0.72	0.18	0.2617	0.0323
		20	69.3	0.67	0.16	0.2416	0.0297
	Poisson	10	84.2	0.09	0.03	0.2575	0.0318
		20	85.7	0.06	0.02	0.2004	0.0245
Poisson	Normal	10	65.7	0.14	0.04	0.0560	0.0068
		20	59.4	0.11	0.03	0.1596	0.0197
	Lognormal	10	87.7	0.05	0.02	0.0673	0.0085
		20	82.7	0.11	0.04	0.0870	0.0111
	Gamma	10	89.6	0.01	0.01	0.0671	0.0085
		20	82.5	0.03	0.01	0.0866	0.0110

Note: The Cushing model and the second set of data that had true *S* values of pink salmon were used in the simulation. Outliers were added when simulating the data. Relative estimate bias (REB) and root mean square error (RMSE) for parameters are shown. GzLM, generalized linear model.

runs with homogeneous residuals was highest when the correct error distributions were used in the GzLM analysis. When the number of outliers increased to 40%, the percentages of homogeneous residuals was no longer highest when the correct error distributions were used in the GzLM analysis. The normal error distribution tended to be more sensitive to outliers compared with the lognormal and gamma distributions. This study shows that even when the GzLM is less effective in identifying the correct error distribution with an increase in the number of outliers, the model error distribution used in the GzLM showing the highest percentage of simulation runs with homogeneous residuals usually pro-

Fig. 3. Relative estimation error (REE) for the two parameters (*a*) ln(α) and (*b*) β in the Cushing model in 1000 simulation runs adding different percentages of outliers in the stock–recruitment data. The number of stock–recruitment observations used in the simulation is 40. The model error distribution used in simulating stock–recruitment data is normal; the model error distributions used in the generalized linear model for the parameter estimation are normal (solid line), lognormal (dashed–dotted line), gamma (dashed line), and Poisson (dotted line).



vides better parameter estimates with smaller REB and RMSE. We present an example (Fig. 3) illustrating the changes in estimation errors for ln(α) and β when different percentages of data were outliers. This suggest that the appropriate model error distribution may not be the underlying true error distribution because of the existence of outliers.

This study suggests that if the model error distribution used in simulating *S–R* data follows the Poisson distribution, then the parameters estimated using different error distributions in the GzLM have smaller differences. Thus, a Poisson-distributed error in *S–R* data is less sensitive to sample size, outliers, and choice of error distributions in the GzLM. The use of the Poisson distribution in the parameter estimation also showed robustness with respect to misspecification of error structure, small sample size, and outliers. For example, when the error distribution used in simulating *S–R* data was normal and the data were contaminated with lognormally distributed outliers, the Poisson error distribution used in the GzLM analysis provided the highest percentage of simulation runs with homogeneous residuals and better parameter estimates. Thus, we recommend using the Poisson distribution in a GzLM analysis of *S–R* data.

The simulated results did not show a clear difference when different *S–R* models were used. Because this is a simulation study and the data sets are based on “true” parameters with random errors, the performance of the models is difficult to compare.

We suggest the GzLM method be used to quantify *S–R* data relationships. The GzLM provides a convenient and effective way to evaluate and identify the appropriate model error distributions for a given set of data and models, leading to improved parameter estimation.

Acknowledgements

We thank Dr. Paul Peng and Dr. Leonard Lye for providing valuable advice. Y. Jiao was supported by a Canadian International Development Agency (CIDA) Marine Science Scholarship. We thank the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Natural Sciences and Engineering Research Council of Canada (NSERC) for providing the major funds for the "Coasts Under Stress" project through the SSHRC Major Collaborative Research Initiatives (MCRI) program.

References

Bajdik, C.D., and Schneider, D.C. 1991. Models of the fish yield from lakes: does the random component matter? Can. J. Fish. Aquat. Sci. **48**: 619–622.

Cadigan, N.G., and Myers, R.A. 2001. A comparison of gamma and lognormal maximum likelihood estimators in a sequential population analysis. Can. J. Fish. Aquat. Sci. **58**: 560–567.

Carroll, R.J., and Ruppert, D. 1984. Power transformations when fitting theoretical models to data. J. Am. Stat. Assoc. **79**: 321–328.

Chen, Y., and Paloheimo, J.E. 1995. A robust regression analysis of recruitment in fisheries. Can. J. Fish. Aquat. Sci. **52**: 993–1006.

Chen, Y., and Paloheimo, J.E. 1998. Can a more realistic model error structure improve the parameter estimation in modeling the dynamics of fish populations? Fish. Res. **38**: 9–17.

Deriso, R.B., Quinn, T.J., and Neal, P.R. 1985. Catch-age analysis with auxiliary information. Can. J. Fish. Aquat. Sci. **42**: 815–824.

Dobson, A.J. 1990. An introduction to generalized linear models. Chapman & Hall, New York.

Fogarty, M.J. 1993. Recruitment distributions revised. Can. J. Fish. Aquat. Sci. **50**: 2723–2728.

Hennemuth, R.C., Palmer, J.E., and Brown, B.E. 1980. A statistical description of recruitment in eighteen selected fish stock. J. Northwest Atl. Fish. Sci. **1**: 101–111.

Hilborn, R., and Walters, C. 1992. Quantitative fisheries stock assessment: choice, dynamics, and uncertainty. Chapman and Hall, New York.

Hinrichsen, R.A. 2001. High variability in spawner–recruitment data hampers learning. Can. J. Fish. Aquat. Sci. **58**: 769–776.

Lindsey, J.K. 1997. Applying generalized linear models. Springer-Verlag, New York.

McCullagh, P., and Nelder, J.A. 1983. Generalized linear models. 1st ed. Chapman and Hall, London.

McCullagh, P., and Nelder, J.A. 1989. Generalized linear models. 2nd ed. Chapman and Hall, London.

Myers, R.A., and Barrowman, N.J. 1996. Is fish recruitment related to spawner abundance? Fish. Bull. **94**: 707–724.

Myers, R.A., Bridson, J., and Barrowman, N.J. 1995. Summary of worldwide spawner and recruitment data. Can. Tech. Rep. Fish. Aquat. Sci. No. 2024.

Myers, R.H., Montgomery, D.C., and Vining, G.G. 2001. Generalized linear models: with applications in engineering and the sciences. John Wiley and Sons, Inc., New York.

Nelder, J.A., and Wedderburn, R.W.M. 1972. Generalized linear models. J. R. Stat. Soc. A, **135**: 370–384.

Power, M. 1996. The testing and selection of recruitment distributions from North Atlantic fish stocks. Fish. Res. **25**: 77–95.

Quinn, T.J., and Deriso, R.B. 1999. Quantitative fish dynamics. Oxford University Press, New York.

Schnute, J.T. 1991. The importance of noise in fish population models. Fish. Res. **11**: 197–223.

Shelton, P.A. 1992. The shape of recruitment distributions. Can. J. Fish. Aquat. Sci. **49**: 1754–1761.

The Math Works Inc. 2002. MATLAB — the language of technical computing. The Math Works Inc., Natick, Mass.

White, G.C., and Bennetts, R.E. 1996. Analysis of frequency count data using the negative binomial distribution. Ecology, **77**: 2549–2557.