available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel

A modelling framework for analysing the reproductive output of individual plants grown in monoculture

C. Brophy[a,*], D.J. Gibson[b], P.M. Wayne[c], J. Connolly[a]

[a] UCD School of Mathematical Sciences, Environmental & Ecological Modelling Group,
University College Dublin, Belfield, Dublin 4, Ireland
[b] Department of Plant Biology, Center for Ecology, Southern Illinois University, Carbondale, IL 62901-6509, USA
[c] Harvard Medical School, Harvard University, Osher Institute, Landmark Ctr Suite 22A, 401 Park Dr, Boston, MA 02215, USA

ARTICLE INFO

Article history:
Received 15 December 2006
Received in revised form
5 April 2007
Accepted 12 April 2007
Published on line 31 May 2007

Keywords:
Reproductive output
Modelling issues
Zero values
Random effects
Truncated regression
Mixture distribution
Pseudoreplication
Sinapis arvensis

ABSTRACT

We propose a framework for modelling plant reproductive data. Several statistical issues can arise in the analysis of reproductive data and this paper develops a framework for dealing with them. The relationship between reproductive output and plant biomass regularly follows a log–log allometric regression. Frequently, a number of plants do not reproduce and the corresponding zero reproductive values do not fall easily within this standard regression framework. Truncated regression allows zero values to be incorporated appropriately in the allometric relationship. We also propose a mixture-model method to deal with outlier values that do not follow the allometric relationship, for example large plants that have zero reproductive output values. Reproductive data from plants grown together in pots or in field plots may not be independent and this dependence should be dealt with in each of the above analyses. We illustrate our method using either generated data or data from an experiment examining reproductive output in *Sinapis arvensis* and provide the programming tools used. We test our methods and compare them with widely used alternatives using simulation studies. These studies validate the use of our proposed approach and show that some of the alternatives produce seriously biased estimates of model parameters. We also present a general graphical aid to assist the selection of the appropriate method to analyse plant reproductive data.

1. Introduction

The relationship between plant reproductive output (R) and plant biomass (DM) is often described by linear allometric regression of log(R) on log(DM) (Harper, 1977; Sugiyama and Bazzaz, 1998; Sletvold, 2002). This method can become problematic when some plants do not produce reproductive structures ($R=0$) and this has led to a number of different strategies for estimating the coefficients of the allometric relationship. Values with $R=0$ have been excluded from the analysis (Sugiyama and Bazzaz, 1998; Sletvold, 2002) or included as zero values, with analysis carried out on untransformed data (Thompson et al., 1991). Schmid et al. (1994) argued that excluding zero responses could lead to biased estimates of the regression coefficients while including the zeros directly in a regression analysis could violate the assumptions of the method. They proposed a truncated regression model to deal with this problem. Truncated regression assumes an allometric relationship between plant reproductive output and size, with zero values arising from an inability

[*] *Corresponding author*. Tel.: +353 1 7167112; fax: +353 1 7161186.
E-mail address: Caroline.Brophy@ucd.ie (C. Brophy).

0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.04.008

to observe reproductive output below a certain threshold level.

Obtaining reproductive biomass frequently requires destruction of the plant, so observing it repeatedly on individual plants is not always possible. When plants are harvested at one time point, it is not unusual to observe a subpopulation of non-reproductive individuals concurrent with reproductive individuals. This can result from spatial heterogeneity in the environment limiting reproduction to individuals in specific, favourable patches (e.g. Gibson et al., 2002), the occurrence of a genetically distinct non-reproductive subpopulation (Wesselingh and de Jong, 1995) or a spread in the timing of reproduction due to co-occurring genetically distinct subpopulations (Rajakaruna et al., 2003) or ontogenetic drift (Evans, 1972; Mc Connaughty and Coleman, 1999).

Plants exhibit allometric growth and size-dependent reproductive allocation (Weiner, 2004). While generally the non-reproductive members of a population are among the smallest (e.g. Vega et al., 2000) they can take a range of sizes. For truncated regression analysis to be successful it is assumed that all plants with zero R and non-zero R follow the same allometric relationship between R and DM. Outliers that clearly do not follow the allometric relationship, for example large plants with zero reproductive output, create a further problem in assessing reproductive output.

The data used in the study of reproductive output is frequently derived from plants growing in clusters (e.g. in pots). Responses from individual plants grown in clusters may be correlated, either due to unexplained differences between pots or as a result of within pot competition (Machin and Sanderson, 1977; Schneider et al., 2006). This type of correlation among responses (hereafter, the 'pot effect') is a form of pseudoreplication and failing to allow for it can cause spuriously significant results from statistical analysis (Hurlbert, 1984). The same concerns arise from the analysis of data collected from plants growing in patches or stands in natural settings. Analysis of the relationships among plant parts for plants harvested at a single time may preclude proper allowance for the effects of ontogenetic drift (Mc Connaughty and Coleman, 1999; Weiner, 2004). However, for plants grown together in pots, the analysis does reflect the effects of neighbour competition on response, part of which may be to spread the trajectories of individual growth and increase the effect of ontogenetic drift.

This paper was motivated by examining data from an experiment on *Sinapis arvensis* (detailed in methodology section). Reproductive biomass and aboveground biomass were recorded at a single point in time and are illustrated on the logarithmic scale in Fig. 1. There were many plants with zero reproductive biomass and a number of them were large plants (circled) suggesting an outlier group from an otherwise strong allometric relationship. Plants in this experiment were grown together in pots at a range of densities and so plants within a pot may be correlated. The purpose of this paper is to present a robust framework that details models for analysing reproductive output where

(a) the data follows a linear allometric relationship,
(b) the data includes plants with zero reproductive output in an allometric relationship,

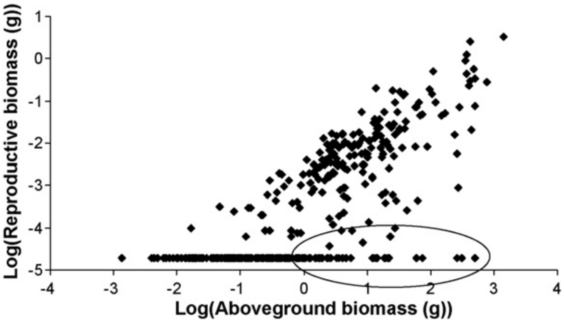

Fig. 1 – The logarithm of reproductive biomass vs. the logarithm of aboveground biomass of individual plants of *Sinapis arvensis*. Approximately, 49% of plants did not reproduce and the logarithm of their reproductive biomass is represented by −4.71. The plot suggests a linear allometric relationship between log(reproductive biomass) and log(aboveground biomass); however the circled values do not support this relationship.

(c) in addition to (b) the data also includes plants with zero reproductive output that are not part of the allometric regression relationship.

We examine each of these for data from spaced plants and also where data come from plants grown together. We develop a novel modelling framework for dealing with these issues in Section 2, and in Section 3 we check our proposed framework against alternatives using either artificially generated data or data from the aforementioned experiment on *S. arvensis*. We use simulation studies to examine the effect of relaxing model assumptions. In Section 4 we provide a generic framework to assist the selection of the appropriate method to analyse reproductive data.

2. Methodology

2.1. *Experiment using S. arvensis*

We will illustrate our methods using data from an experiment on *S. arvensis* that contains many of the issues that commonly arise with reproductive output data. *S. arvensis* L. (formerly *Brassica kaber* var. *pinnafitida* (Stokes) L. C. Wheeler) (field mustard, charlock, Brassicaceae), an annual species native to Eurasia (Fogg, 1950) is an important agricultural weed in the mid-western regions of North America (Gleason and Cronquist, 1991; Warwick et al., 2000). On December 23, 1996, seed were sown directly into 5.5 L 25 cm diameter round pots (stands) in a greenhouse at Harvard University. Seeds were sown at six densities: 1, 2, 4, 8, 16 and 32 plants pot^{-1}. There were three blocks each containing 14 pots. The pots were harvested on February 17, 1997 when a large number of flowers had matured into fruits but before many leaves had senesced (LAI = 2.9). After separating leaves and support structures (stems and petioles), aboveground biomass for individual plants was oven dried at 70 °C for one week and weighed. The biomass of all reproductive structures (flowers and fruits) of all individuals was also measured. In total there were 42 pots and 374 measurable plants. Further details on the experiment are in Wayne et al. (1999). We define R to now be reproductive biomass, DM to be aboveground plant biomass and a prefix of L indicates natural logarithm.

An initial plot of the data showed a strong allometric relationship between LR and LDM for reproducing plants (Fig. 1). Many plants did not reproduce (ca. 49%) and a number of these were plants with large DM values (circled) clearly were not part of this allometric relationship.

2.2. *The modelling framework*

This section details a theoretical framework for the analysis of reproductive output from plants. We implement all components of this framework using generated data (see supplementary online material) and illustrate the analysis of the most complex model using data from the *S. arvensis* experiment in Section 2.1.

2.2.1. *Linear allometric regression*

The relationship between reproductive biomass (R) and size (DM) is usually described by log–log linear allometric regression. For the ith plant the model is of the form:

$$\mathrm{LR_i} = \beta_0 + \beta_1'\mathrm{LDM_i} + \varepsilon_\mathrm{i} \qquad (1)$$

where β_0 is the intercept, $\beta_1^{'}$ is the slope of the allometric relationship between LR and LDM and ε_i is the residual term which is assumed to be normally distributed with mean zero and variance σ^2 and independent of other residual terms. If $\beta_1^{'} = 1$, the plant allocates a constant proportion of biomass (DM) to reproduction regardless of size, while $\beta_1^{'} > 1\,(< 1)$ indicates that larger plants allocate a higher (lower) proportion of biomass to reproduction than smaller plants. Additional variables, for example a treatment or block effect, may also be included in this model.

Reproductive biomass may also be scaled by aboveground biomass to give the ratio = R/DM, which we define to be reproductive allocation (RA). Reproductive allocation has been defined in other ways previously, for example the ratio of reproductive biomass to total (above and below ground) biomass (e.g. He et al., 2005), and the ratio of reproductive biomass to vegetative biomass (e.g. Huxman et al., 1999). The following methods can be easily modified to deal with these alternative definitions of RA. Using RA as the response, model (1) becomes:

$$\mathrm{LRA_i} = \beta_0 + \beta_1\mathrm{LDM_i} + \varepsilon_\mathrm{i} \qquad (2)$$

where β_0, ε_i are as defined above and $\beta_1 = \beta_1^{'} - 1$ is the slope of the allometric relationship between LRA and LDM. If β_1 is zero, the plant allocates a constant proportion to reproduction regardless of size. A positive (negative) β_1 indicates that larger plants allocate a higher (lower) proportion to reproduction than smaller plants. Models (1) and (2) can be fitted using ordinary linear regression (Sokal and Rohlf, 1995) (see supplementary online material program 1).

Reproductive data frequently comes from experiments where plants are grown in pots in a greenhouse or in stands in the field. In this paper we will refer to the clustering of plants in pots but the applications apply to both cases. Plants within a pot are potentially correlated and model (2) can be extended to include a random pot effect that induces this correlation. For the ith plant from the jth pot:

$$\mathrm{LRA_{ij}} = \beta_0 + \beta_1\mathrm{LDM_{ij}} + u_j + \varepsilon_{ij} \qquad (3)$$

where β_0 and β_1 are as described for (2) and ε_{ij} is the random term for the ith plant from the jth pot and u_j is the random effect for the jth pot; ε_{ij} and u_j are assumed normally independently distributed with mean zero and variance σ_1^2 and σ_2^2, respectively and independent of each other. Model (3) is fitted using mixed model software (see supplementary online material program 2). A positive correlation is induced on plants growing in the same pot and while this may be offset somewhat by a negative within pot competition, model (3) assumes the resultant correlation is still positive.

Statistical theory states that if model (2) is fitted when model (3) describes the true relationship between LRA and LDM the parameter estimates of β_0 and β_1 are unbiased but their standard errors are biased. We tested the effect of omitting the random pot effect when model (3) is appropriate on the standard errors of the estimates of β_0 and β_1 using a simulation study. We simulated a theoretical relationship between LRA and LDM with two sources of variability according to model (3) with $\beta_0 = -3.14$ and $\beta_1 = -0.11$ giving the relationship: $\mathrm{LRA_{ij}} = -3.14–0.11\mathrm{LDM_{ij}} + u_j + \varepsilon_{ij}$. Using the 374 LDM values from the experiment described in Section 2.1 we generated LRA values assuming ε_{ij} and u_j to be normally distributed with mean 0 and known variances σ_1^2 and σ_2^2, respectively. We generated 1000 datasets of size 374 for each combination of $\sigma_1^2 = 0.416,\ 0.523$ and 0.658 and $\sigma_2^2 = 0, 0.01, 0.02, 0.05, 0.25, 0.45$ and 2. The parameter values chosen here and for all simulation studies described in this paper are motivated by analysis (described in full later) on data from the *S. arvensis* experiment (Section 2.1). We fitted the two allometric regression models, (2) and (3), to each dataset and calculated the average standard error for the slope and intercept for each set of 1000 for each model. We compared the two models using the ratio of the average standard errors for the model without (2) to that with (3) the random effect.

2.2.2. *Modelling data with zero RA values present*

The modelling methods in Section 2.2.1 are in question when a number of zero RA values are present. Methods that have been employed previously include ignoring zeros (Sugiyama and Bazzaz, 1998; Sletvold, 2002) and leaving them in as zeros and analysing the untransformed data (Thompson et al., 1991). Schmid et al. (1994) proposed truncated allometric regression as a method to deal with zeros. Truncated regression assumes an allometric relationship between plant RA and size, with responses for which RA > 0 treated as in usual allometric regression (Section 2.2.1) and zero RA values assumed to arise from an inability to observe reproductive output below a certain threshold level, i.e. the true RA value has been truncated. A truncated regression model can be fitted using the method of maximum likelihood. Assuming the data is clustered in pots, if the ith plant from the jth pot has $\mathrm{RA}_{ij} > 0$, then its contribution to the likelihood is

$$f(\mathrm{LRA}_{ij}) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{1}{2}\frac{(\mathrm{LRA}_{ij}-\beta_0-\beta_1\mathrm{LDM}_{ij}-u_j)^2}{\sigma_1^2}\right) \qquad (4)$$

i.e. f is the likelihood of LRA_{ij} based on a regression model. If the ith plant from the jth pot has $\mathrm{RA}_{ij} = 0$, its contribution to the likelihood is

$$F(\mathrm{LRA}_{ij}) = \int_{-\infty}^{\lambda-\mathrm{LDM}_{ij}} \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{1}{2}\frac{(LRA_{ij}-\beta_0-\beta_1 LDM_{ij}-u_j)^2}{\sigma_1^2}\right) \times dLRA_{ij} \qquad (5)$$

i.e. F is the likelihood of LRA_{ij} based on assuming plants with zero RA are caused by an inability to observe R below a certain threshold but that they follow the same allometric relationship as those plants with $\mathrm{RA} > 0$. This integral represents the probability of a plant of a given size from a given pot having reproductive biomass below the observable threshold (equal to the shaded area in Fig. 2). Note that calculating LRA_{ij} for plants with $\mathrm{RA} = 0$ is problematic since $\log(0) = -\infty$. To deal with this issue, when $R_{ij} = 0$, we let $\mathrm{LR}_{ij} = \lambda$, where λ is a small value less than or equal to the logarithm of the smallest observed value of R_{ij}. Hence, when $\mathrm{RA}_{ij} = 0$, LRA_{ij} is $\lambda - \mathrm{LDM}_{ij}$. The likelihood of all the data is

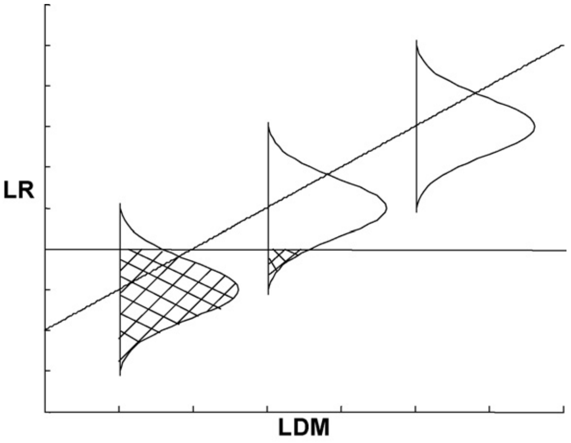

Fig. 2 – Hypothetical allometric relationship between LR (log(reproductive biomass)) and LDM (log(aboveground biomass)). The horizontal line indicates the threshold below which reproductive biomass will not be observed. The bell-shaped curves represent the normal distribution assumed around the expected mean LR value for three given LDM values. With truncated regression, plants with zero and with non-zero reproductive biomass are assumed to follow the same relationship. The probability of zero reproductive biomass for a plant of a given size corresponds to the shaded area (i.e. the probability of R being below the observable threshold). For this hypothetical positive relationship between LR and LDM the smaller the plant size, the higher this probability becomes (cf. the two shaded areas). When truncated regression is used to estimate an allometric relationship, non-zero values are treated as in normal linear regression (using Eq. (4)) and for zero values, it considers the probability equal to the shaded area (the integral in (5)).

$$\text{likelihood} = \prod_{\mathrm{RA}_{ij}>0} f(\mathrm{LRA}_{ij}) \prod_{\mathrm{RA}_{ij}=0} F(\mathrm{LRA}_{ij}) \qquad (6)$$

This model can be fitted using non linear regression software that allows for random effects. We used the NLMIXED procedure in the software SAS version 9.1 (SAS Institute Inc., Cary, NC, USA; see supplementary online material program 3). Note that when there are no zero values and model (3) is fitted using maximum likelihood, the contribution of each individual to the likelihood is in Eq. (4).

We use simulation to examine the truncated regression method and two alternatives as approaches to dealing with zero RA values. Ignoring plants with no reproductive output when determining the relationship between size and reproductive output (Sugiyama and Bazzaz, 1998; Sletvold, 2002) has been advised against (Schmid et al., 1994; Underwood, 1997; Vega et al., 2000; Gibson, 2002; Quinn and Keough, 2002). However, this approach and the method of including plants as zero values and analysing on the untransformed scale (Thompson et al., 1991) may be satisfactory when only a small number of zeros are present. We test these approaches and the truncated regression method by fitting three models to simulated truncated datasets: (i) an allometric model to values above the truncation point only (the threshold at which it is assumed reproductive biomass cannot be observed below); (ii) an allometric model with truncated values set equal to the truncation value on the log scale (this is analogous to including zeros and analysing on untransformed data); and (iii) a truncated allometric regression model. We simulated a theoretical relationship between LRA and LDM defined by model (2), $\mathrm{LRA}_i = \beta_0 + \beta_1\mathrm{LDM}_i + \varepsilon_i$, for a range of known values of β_0 and β_1. We assumed that the ε_i values are normally independently distributed with mean zero and known variance σ^2. Using the 374 LDM values from the *S. arvensis* experiment (Section 2.1) we generated 1000 datasets of size 374 at each combination of $\beta_0 = -3.341$, -3.14 and -2.946, $\beta_1 = -0.2$, -0.1 and 0 and $\sigma^2 = 0.416$, 0.523 and 0.658. We then took two subsets of size 100 and 200 at random from the original 374 values and repeated this generating process. We truncated each dataset at six truncation levels of increasing severity based on the assumption that reproductive mass below a certain value could not be observed. The 0%, 1%, 5%, 10%, 30% and 50% percentiles of R (=reproductive biomass) were calculated for each dataset and all values below the percentile were considered unobservable and were truncated by being set equal to the percentile value. LRA was recalculated for truncated values. We fitted the three models to each generated dataset and calculated the bias in the estimated values of β_0, β_1 and σ^2. Bias in the estimation of β_0 and β_1 was calculated in units of their true standard error. The true standard error was assumed to be the standard deviation over all parameter estimates for model (3). We calculated the bias as the parameter estimate minus the true parameter value divided by the true standard error. Bias for the variance was estimated using the ratio of the estimated variance to the true variance.

We extended the simulation study in Section 2.2.1 to examine the effect of ignoring a random pot effect for a truncated regression model. Using the truncation method described in the previous paragraph, we truncated each generated dataset from the simulated datasets in Section 2.2.1. We examined whether increasing severity of truncation affected the precision with which parameters where estimated when the random pot effect was ignored.

2.2.3. *Modelling RA values that do not follow the allometric relationship*

Truncated regression assumes that zero reproductive biomass arises from an inability to observe reproduction below a certain threshold and that all plants (including those with zero and non-zero reproductive biomass) follow the same allometric relationship (Schmid et al., 1994). There may be situations where not all non-reproducing plants follow the same relationship, e.g. the zero reproductive biomass values observed from large plants circled in Fig. 1. This figure suggests a strong allometric relationship but also a group of outlier non-reproducing plants. We propose assuming two groups within the data. We assume the first group follows the allometric relationship and contains all reproducing plants and some non-reproducing plants. We assume the second group does not follow the allometric relationship and contains the remaining non-reproducing plants. We emphasise that plants are divided depending on whether or not an individual follows the allometric relationship and not on its reproductive status so that while group 1 contains all reproducing plants, it also contains some non-reproducing plants. An individual non-reproducing plant cannot be unambiguously assigned to one or other group; we say it is in group two with probability *p*, and in group one with probability 1−*p*, where *p* has to be estimated from the data. We expand the likelihood in (6) to include these two probabilistic groups to give a finite mixture model (Mc Lachlann and Peel, 2000). Where LRA_{ij} represents the log reproductive allocation of the *i*th plant from the *j*th pot, the likelihood of all the data is:

$$\text{likelihood} = \prod_{RA_{ij}>0} (1-p) f(LRA_{ij}) \prod_{RA_{ij}=0} [(1-p)F(LRA_{ij}) + p] \qquad (7)$$

where *f* and *F* are as described in (4) and (5), respectively; *p* is the probability that an individual LRA value does not follow the allometric relationship; *p* can be a constant or can depend on explanatory variables such as size. We assume the allometric relationship is as in (3) i.e. linear in LDM and with a random effect, so *f* and *F* are defined as before. Model (7) can be fitted using the EM algorithm or non-linear regression software that allows for a random effect. We used the NLMIXED procedure in SAS (see supplementary online material program 3). In modelling *p*, we use the logit transformation to constrain the estimates of *p* to lie between 0 and 1 (Collett, 1993) and include a random pot effect to account for potential correlation between plants within a pot. One possible form for the model for *p* is:

$$\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 \mathrm{LDM} + w_j \qquad (8)$$

where α_0 is a constant, w_j is a normally distributed random pot effect assumed to have mean zero and variance σ_3^2 and its covariance with u_j is γ.

The many large plants with zero RA in Fig. 1 strongly support the concept of the mixture model defined in the previous paragraph. If this second group did not exist, would our model provide evidence for it? To answer this question we simulated data to check the allocation of zero responses to a putative second group in situations where one did not exist. We simulated a theoretical relationship between LRA and LDM defined by model (2) with $\beta_0 = -3.14$ and $\beta_1 = -0.11$, giving the relationship: $\mathrm{LRA}_i = -3.14–0.11\mathrm{LDM}_i + \varepsilon_i$. We assumed ε_i to be normally distributed with mean zero and known variance σ^2. Using the 374 LDM values from the data from the *S. arvensis* experiment, we generated 1000 datasets at each of $\sigma^2 = 0.402$, 0.468, and 0.545. We truncated each dataset as described above in Section 2.2.2 and fitted the mixture model defined by the likelihood in (7) to each dataset with *p* assumed constant and omitting the random pot effect. We examined the predicted proportions in group two for each model.

2.2.4. *A test using the S. arvensis dataset*

The *S. arvensis* dataset described in Section 2.1 contains all the issues we have discussed in this section and so we fitted the mixture model defined by the likelihood in (7) to it. We maximised the log of the likelihood function using the NLMIXED procedure using the software SAS version 9.1 (SAS Institute Inc., Cary, NC, USA) to obtain estimates of all parameters of the two component models defined by (3) and (8). The blocking in the design was not important and so was omitted in analysis. The models can be readily extended to examine multiple factors simultaneously.

We predicted RA conditional on being in group 1 and predicted the probability of being in group 2 from our models. Predictions from the two components of the mixture model were then combined to give the joint effects of DM on RA but standard errors for these predictions or the effect observed were not readily available. We obtained standard errors using a bootstrap analysis (Efron and Tibshirani, 1993) as follows. A thousand bootstrap datasets were constructed by re-sampling with replacement at the pot level within each density and again at the plant level within pot and the model was fitted for each of these samples. The standard error for any prediction from the original model is calculated as the standard deviation of the predictions obtained from these 1000 models.

3. Results

3.1. *Results from simulation studies*

3.1.1. *The precision of parameter estimates in an allometric regression model when a random pot effect is ignored*

Our simulation study showed that ignoring correlated pot responses in a linear allometric relationship causes the precision with which the intercept and slope are estimated to be affected (Table 1; row for 0% truncation). Standard errors for the intercept were always underestimated when the random pot effect was ignored (indicated by values always less

Table 1 – Results from a simulation study of the standard errors of parameter estimates when potential dependence among plants grown together in pots is ignored

Truncation severity (%)	Intercept σ_2^2(between pot variance)							Slope σ_2^2 (between pot variance)						
	0	0.01	0.02	0.05	0.25	0.45	2	0	0.01	0.02	0.05	0.25	0.45	2
0	0.95	0.87	0.79	0.65	0.44	0.40	0.35	0.99	0.97	0.95	0.92	0.93	1.00	1.38
1	0.95	0.87	0.79	0.65	0.44	0.40	0.35	0.99	0.97	0.95	0.92	0.93	1.00	1.38
5	0.95	0.87	0.79	0.65	0.44	0.40	0.35	0.99	0.97	0.95	0.92	0.93	1.00	1.37
10	0.95	0.87	0.80	0.66	0.45	0.40	0.35	0.99	0.97	0.95	0.92	0.93	0.99	1.35
30	0.95	0.89	0.83	0.70	0.48	0.44	0.38	0.99	0.97	0.96	0.93	0.92	0.97	1.27
50	0.96	0.92	0.88	0.78	0.57	0.51	0.44	0.99	0.98	0.97	0.94	0.92	0.95	1.19

The LRA values were simulated according to the relationship: $\text{LRA}_{ij} = -3.14\text{–}0.11\text{LDM}_{ij} + u_j + \varepsilon_{ij}$ where u_j (a random pot effect for the jth pot) and ε_{ij} (a random effect for the ith plant from the jth pot) are assumed normally independently distributed with mean zero known variances σ_2^2 (pot-to-pot variance) and σ_1^2 (within pot variance), respectively. One thousand datasets were generated for each combination of a range of truncation severities and values of between and within pot variances. The average standard error was calculated for the intercept and slope for models with and without a random pot effect and the ratio (without/with) is presented. Results are presented only for $\sigma_1^2 = 0.468$ as there was very little change over the range of σ_1^2 values. Values of < 1, 1 and > 1 indicate too small a standard error, correct standard error and too large a standard error, respectively, when the random pot effect is ignored.

than 1) and the extent of the underestimation increased as the pot-to-pot variance (σ_2^2) increased. When the residual variance and the pot-to-pot variance were of similar magnitude ($\sigma_2^2 = 0.45$), ignoring the random effect caused a 60% underestimation in the standard error for the intercept. Standard errors for the slope were usually underestimated but when σ_2^2 was larger than σ_1^2 the standard errors were overestimated. When the pot-to-pot variance was much larger than the residual variance ($\sigma_2^2 = 2$), the standard errors for the slope were overestimated by almost 40%. When there is no pot-to-pot variation ($\sigma_2^2 = 0$) the ratio deviates from 1 for both the intercept and the slope showing a penalty from fitting a parameter that is not necessary.

3.1.2. *Truncated regression model testing*

Bias in the estimates of the intercept and slope (measured in units of true standard error of the parameter) in models (i) truncated values omitted and (ii) truncated values included equal to the truncation level, was severe even at low levels of truncation but was negligible in model (iii) the truncated regression model (Table 2). Ignoring 10% truncated values (model (i)) of a dataset of size 200 caused an overestimation in the intercept of just over two-fold the standard error and an underestimation of the slope by nearly two and a half times the standard error. While including 10% truncated values at the truncation value (model (ii)) of a dataset of size 200 caused an overestimation in the intercept of just over one standard error and an underestimation of the slope by just over one and a half times the standard error. The bias worsened for both parameters and both models as sample size and truncation severity increased. The variance parameter (the ratio of estimated variance to true variance) was underestimated severely in models (i) and (ii) at higher truncation levels and was slightly underestimated in model (iii). Ignoring 10% of values in a dataset of size 200 caused the variance to be underestimated by 12% while ignoring 50% of values caused an underestimation of 34%. Including truncated values at the truncation value caused an even stronger underestimation of the variance; 10% and 50% of truncated values caused 13% and 46% underestimation, respectively.

3.1.3. *The precision of parameter estimates in a truncated regression model when a random pot effect is ignored*

Ignoring a random effect causes problems with the precision with which parameters are estimated when using truncated regression. The degree of the problem does not worsen much with increasing truncation severity (Table 1) and so the results described for the 0% truncation level (Section 3.1.1) still hold.

3.1.4. *Tests for p when a second group does not exist*

Our simulation study showed that if a second group is not present the mixture modelling approach outlined is unlikely to spuriously suggest strong evidence for a second group (Table 3). When the truncation severity was low the estimate of p was low and as truncation severity increased, the estimate of p also increased; for example with $n = 100$ and $\sigma^2 = 0.402$, p was estimated to be 0.0006 and 0.0115 at the 1% and 50% truncation severities, respectively.

3.2. *Results from fitting the mixture model to the S. arvensis dataset*

The fitted mixture model has two components: the allometric regression model for group 1 (all reproducing and some non-reproducing plants) is:

$$L\hat{R}A_{ij} = -3.04 - 0.08\text{LDM}_{ij} \tag{9}$$

where the symbol ˆover LRA indicates predicted; and the model for p, the probability of being in the second group (remaining non-reproducing plants) is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.21 - 1.47\text{LDM} \tag{10}$$

Two components were necessary in the model; tested using the BIC statistic (Schwartz, 1978; Mc Lachlann and Peel, 2000). The standard errors of parameter estimates and likelihood ratio tests of significance for parameters and all variance component estimates are shown in Table 4. Inclusion of the

Table 2 – Evaluation, for simulated truncated data, of bias in parameter estimates of the slope, intercept and variance of (i) a linear regression model ignoring truncated responses (ii) a linear regression model replacing truncated responses by the truncation value and (iii) the truncated regression model

Parameter	Truncation severity (%)	Model								
		(i) Sample size			(ii) Sample size			(iii) Sample size		
		100	200	374	100	200	374	100	200	374
Intercept	1	0.2	0.3	0.3	0.0	0.1	0.1	−0.01	−0.02	0.00
	5	0.8	1.1	1.6	0.3	0.5	0.7	−0.02	−0.02	0.00
	10	1.4	2.1	3.0	0.8	1.1	1.6	−0.02	−0.02	0.00
	30	3.7	5.3	7.4	3.0	4.4	6.1	−0.02	−0.02	0.02
	50	5.2	7.2	10.0	4.8	7.0	9.6	−0.01	−0.01	−0.01
Slope	1	−0.3	−0.4	−0.5	−0.1	−0.1	−0.1	0.03	0.01	0.00
	5	−1.0	−1.4	−1.9	−0.5	−0.7	−0.9	0.03	0.01	0.00
	10	−1.6	−2.3	−3.1	−1.1	−1.6	−2.1	0.03	0.01	−0.01
	30	−3.0	−4.3	−5.7	−3.3	−4.8	−6.3	0.01	0.00	−0.02
	50	−3.4	−4.7	−6.4	−4.7	−6.7	−9.0	0.01	0.00	0.00
Variance	1	0.96	0.96	0.97	0.97	0.98	0.98	0.99	0.99	0.99
	5	0.91	0.91	0.92	0.92	0.93	0.93	0.99	0.99	0.99
	10	0.87	0.88	0.88	0.87	0.87	0.87	0.99	0.99	0.99
	30	0.76	0.78	0.78	0.70	0.71	0.70	0.98	0.99	0.99
	50	0.65	0.66	0.66	0.53	0.54	0.53	0.98	0.99	0.99

Bias for each of the three parameters was computed for each truncation severity by sample size combination. Bias for the intercept and slope were calculated by the parameter estimate minus the true parameter value relative to the true standard deviation. Unbiased intercept and slope are indicated by bias ≈0. A positive (negative) value indicates overestimation (underestimation) of the parameter estimate, for example, a value of −2 (0.5) indicates the parameter is underestimated by two-times (overestimated by half) the standard error. The units of measurement are bias relative to the true standard deviation of the parameter estimate. Bias for the variance was estimated using the ratio of the estimated variance to the true variance. Unbiased variance is indicated by bias ≈1, otherwise the values are interpreted as: e.g. a value of 0.5 (2) indicates that the estimate of the variance is half (twice) the size of the true variance.

Table 3 – Estimation of *p*, the proportion in group 2, from the full model, defined by the likelihood in (7), when *p* is actually zero

Sample size	σ^2	Truncation severity						
		1%	5%	10%	20%	30%	40%	50%
100	0.402	0.0006	0.0009	0.0010	0.0019	0.0033	0.0056	0.0115
100	0.545	0.0006	0.0012	0.0016	0.0034	0.0062	0.0111	0.0172
200	0.402	0.0003	0.0004	0.0005	0.0010	0.0021	0.0036	0.0054
200	0.545	0.0003	0.0006	0.0010	0.0021	0.0044	0.0062	0.0091
374	0.402	0.0001	0.0002	0.0004	0.0007	0.0013	0.0022	0.0041
374	0.545	0.0001	0.0003	0.0006	0.0014	0.0027	0.0040	0.0061

Shown are the average estimate and standard error of p for data simulated for combinations of n, σ^2 (residual variance) and truncation severity. Low estimates of p indicate that the model is unlikely to suggest a second group in a situation where one does not exist.

random terms for pot (σ_2^2 and σ_3^2), tested using a likelihood ratio test, were a necessary feature of the model ($p = 0.007$, 2 d.f.) indicating a positive within plot correlation between plant reproductive responses. The covariance between the two random effects was not necessary ($p = 0.13$) and so was set to 0. The maximum likelihood estimate of λ was −4.1075, the logarithm of the smallest non-zero value of R_{ij}.

For plants in group 1, RA is predicted to decrease with increasing plant size, to about 4.7% for plants with DM = 9 g (Fig. 3(a)). The predicted probability of being in the second group decreases rapidly with plant size (Fig. 3(b)). Overall RA is assessed in Fig. 3(c): the predicted RA for a plant of a given size increases rapidly for small plants and then plateaus at about 4.2% for plants greater than 4 g. Standard errors for predictions from the bootstrap analysis are included in Fig 3(c). Using the bootstrap method, we found effects of DM on RA for plants below 3 g; RA differed at DM = 1 and 2, at DM = 1 and 3 and at DM = 2 and 3 ($p < 0.01$ for each test). There was no effect of DM above 3 g.

4. Discussion

We provide a powerful framework for the analysis of reproductive output in plants. We test our approach against a range of alternatives using simulation studies. We provide a practical graphical aid to help determine the appropriate analysis for reproductive data (Fig. 4). Any reader wishing to apply

Table 4 – Parameter estimates for the final model fitted from the likelihood in (7), with standard errors and significance levels from likelihood ratio tests

	Parameter	Estimate	S.E.	*p*-value
Group 1	Intercept	−3.14	0.099	–
	LDM	−0.11	0.073	0.157
Group 2	Intercept	0.26	0.292	–
	LDM	−1.80	0.237	<0.001
Variance components	σ_1^2	0.523		
	σ_2^2	0.010		
	σ_3^2	0.778		

Also shown are the variance component estimates; σ_1^2 is the within pot variance in group 1; σ_2^2 is the variance of the random effect in group 1; σ_3^2 is the variance of the random effect in group 2. LDM is log(aboveground biomass).

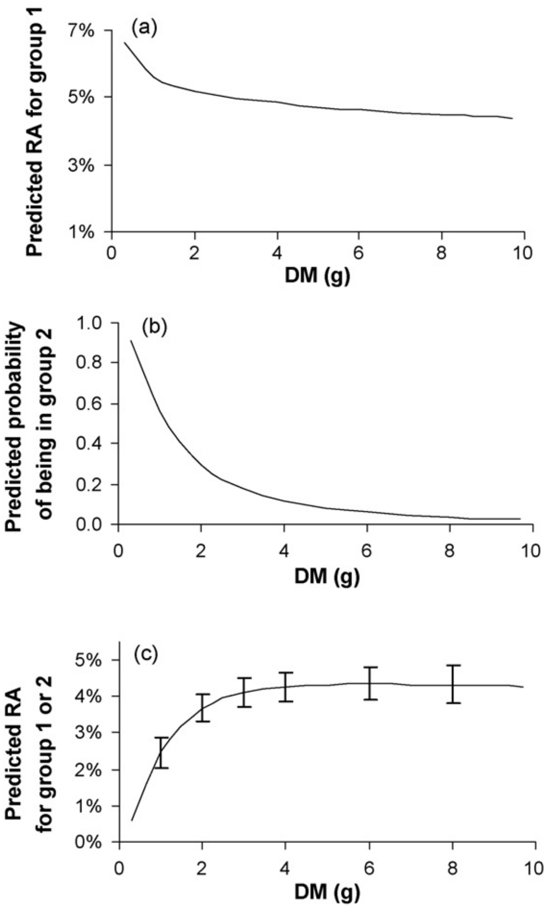

Fig. 3 – The relationship between plant size (DM); and (a) the predicted average RA (%) for an individual in group 1, (b) the predicted probability of an individual being from group 2, and (c) the predicted RA (%) for an individual of given size. (c) also includes ±standard error bars for predicted RA at DM = 1, 2, 3, 4, 6 and 8 g calculated from a bootstrap analysis.

our methods can determine which model is the most suitable for their particular data using this aid. When the appropriate model has been selected the reader can use the SAS programs provided in our supplementary online material to implement all the methods.

Our results support the view that ignoring large numbers of zero RA values is not satisfactory (Schmid et al., 1994; Underwood, 1997; Vega et al., 2000; Gibson, 2002; Quinn and Keough, 2002). Sletvold (2002) and Sugiyama and Bazzaz (1998) dealt with a small number of zero values by omitting them, however, even ignoring 5% of values can cause bias in parameter estimation; our simulation study showed bias in the estimates of slope and intercept to be 1.6 and −1.9 times their standard errors, respectively, at the 5% truncation level with sample size 374 (Table 2). Including zero values directly as zeros in allometric regression (Thompson et al., 1991) also leads to biased parameter estimates; our simulation study showed bias in the estimates of the slope and intercept to be 0.7 and −0.9 times their standard errors at the 5% truncation level with $n = 374$ (Table 2). Bias in both parameter estimates increased with increasing truncation severity. The normality and linearity assumptions in linear regression are also in doubt using this method. Truncated regression assumes the linear relationship continues below the threshold point but including truncated values directly in simple linear regression provides no facility to acknowledge this. Close to the truncation point, the distribution assumed around the expected mean LRA becomes skewed in simple linear regression violating the normality assumption. Mendez and Karlsson (2004) proposed dealing with zero values by supplementing a linear regression analysis for the non-zero RA data with a separate analysis investigating the size dependence of the probability of reproducing. Truncated regression analysis combines these two analyses to give precise estimates of the parameters describing the allometric relationship. When using the two separate analyses, there is no facility to combine results and so the regression line for only reproducing values can mislead as parameter estimates are biased as is shown by our simulation study ignoring truncated values (results for model (i) Table 2). We have used Type I regression throughout. We simulated the effect on our estimates of using Type I estimates when a Type II approach was more appropriate (5% error in measuring DM) and concluded that that our model was not sensitive to up to 5% measurement error in DM (see supplementary online material 4).

Hurlbert (1984) highlighted the dangers of ignoring pseudoreplication in analysis of experimental data. When plants are grown in competition in pots, plants within a pot are not statistically independent: an example of 'sacrificial pseudoreplication' (*sensu* Hurlbert, 1984). Using a simulation study, we verified that the problems associated with ignoring this type of pseudoreplication exist for the allometric regression model (2). We also verified that this problem is present when using the truncated regression model by increasing the apparent precision in estimating the intercept and having a complex effect in respect of the precision with which the slope is estimated (Table 1). We concur with Hurlbert (1984) and others (Underwood, 1997; Gibson, 2002; Quinn and Keough, 2002) that correlated responses should not be ignored in analysis and this applies when implementing any part of our framework to

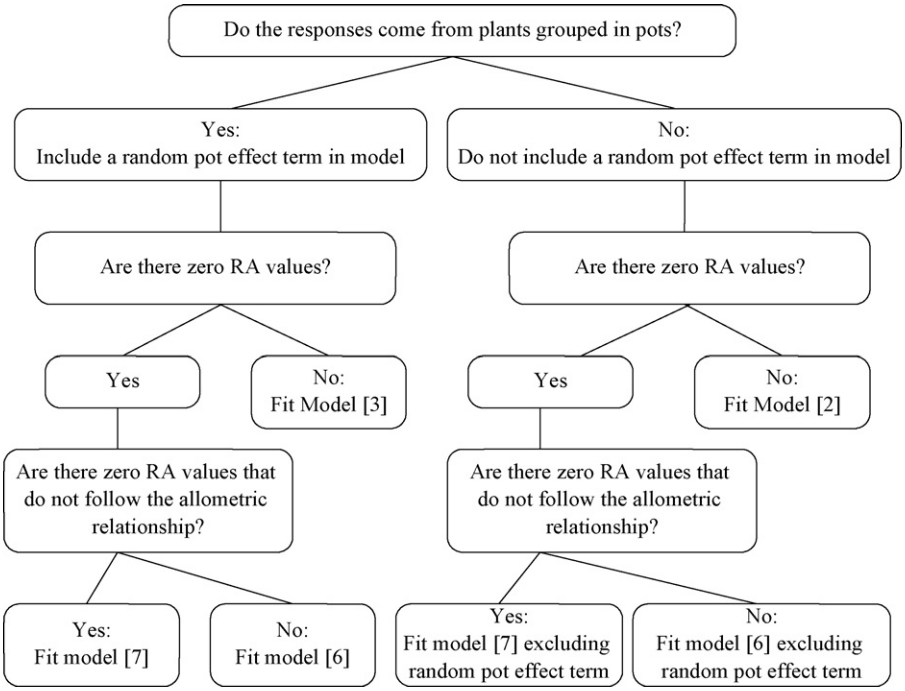
Do the responses come from plants grouped in pots?
Yes:
Include a random pot effect term in model
No:
Do not include a random pot effect term in model
Are there zero RA values?
Are there zero RA values?
Yes
No:
Fit Model [3]
Yes
No:
Fit Model [2]
Are there zero RA values that do not follow the allometric relationship?
Are there zero RA values that do not follow the allometric relationship?
Yes:
Fit model [7]
No:
Fit model [6]
Yes:
Fit model [7] excluding random pot effect term
No:
Fit model [6] excluding random pot effect term

Fig. 4 – Flow chart describing the process of following a logical series of questions to arrive at a suitable model to estimate the relationship between reproductive output and plant size. Each model is described in full in the methods section and code for implementation is available in the supplementary online material.

data from plants grown in pots. Our models allow for this by including a random pot effect (models (3), (6), (7)).

In a truncated allometric relationship the proportion of individuals not reproducing is statistically related to the variables in the model and the normality assumptions of the model; small individuals have a lower probability of having a reproductive mass than do larger individuals. The relative proportions are determined by a normal integral from minus infinity to an upper bound dependent on plant size. Schmid et al. (1994) suggested that in some individuals reproductive material may have been present but not observable at their stage of development and that this was more likely with smaller plants. Membership of the second group of large non-reproducing plants observed in the S. *arvensis* experiment (Fig. 2) is more problematic to explain. It is possible that within the original field collection of S. *arvensis* seed there were two genetically distinct groups that differ in their ability to reproduce under our range of experimental conditions; two seedling cohorts of S. *arvensis* have been observed previously in the field (Fogg, 1950; Edwards, 1980). An alternative explanation is that since all plants were harvested at a single point in time they were at different ontogenetic stages (Mc Connaughty and Coleman, 1999), (see Stanton et al. (2000) for environmental stress-induced ontogenetic shift in S. *arvensis*) giving the large range of plant sizes within the non-reproducing plants. Nevertheless, the robustness of the allometric relationships between reproductive allocation and biomass observed here reinforce the need to consider the appropriate framework for analyzing reproductive output.

The non-significant effect of size on RA in group 1 (Fig. 3(a)) agrees with the findings of Cheplick (2005) who found no relationship between vegetative biomass and RA in general for annuals. However, interpretation of the effect of biomass on RA cannot be judged from Fig. 3(a) alone as the true results are from the combined effect of size from the two groups on RA. From Fig. 3(c) we determine that size does affect RA for small plants, which is in conflict with the literature but for plants bigger than 3 g, there is no effect of size on RA agreeing with the literature. The ability to disaggregate the results into two groups can facilitate further understanding of underlying biological mechanisms.

Acknowledgements

CB is the holder of an Environmental Protection Agency (Ireland) Doctoral Scholarship, JC has received support from an Enterprise Ireland International Collaboration grant for this work, and DG received support from the United States Department of Agriculture. We acknowledge the helpful comments of Ed Reekie, Adrian Dunne, Kristina Stinson and Bodil Frankow-Lindberg.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2007.04.008.

REFERENCES

Cheplick, G.P., 2005. The Allometry of Reproductive Allocation. In: Reekie, E.G., Bazzaz, F.A. (Eds.), Reproductive Allocation in Plants. Elsevier Inc., Amsterdam, pp. 94–125.

Collett, D., 1993. Modelling Binary Data: CRC Texts in Statistical Science. Chapman & Hall/CRC.

Edwards, M., 1980. Aspects of the population ecology of charlock. J. Appl. Ecol. 17, 151–171.

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Chapman & Hall.

Evans, G.C., 1972. The Quantitative Analysis of Plant Growth. University of California Press, Berkeley, CA.

Fogg, G.E., 1950. *Sinapis arvensis* L. J. Ecol. 38, 415–429.

Gibson, D.J., 2002. Methods in Comparative Plant Population Ecology. Oxford University Press, Oxford.

Gibson, D.J., Spyreas, G., Benedict, J., 2002. Life history of *Microstegium vimineum* (Poaceae), an invasive grass in southern Illinois. J. Torrey Bot. Soc. 129, 207–219.

Gleason, H.A., Cronquist, A., 1991. Manual of Vascular Plants of Northeastern United States and Adjacent Canada, second ed. New York Botanical Garden, Bronx, NY 10458 USA, p. 910.

Harper, J.L., 1977. Population Biology of Plants. Academic Press, London.

He, J.S., Wolfe-Bellin, K.S., Bazzaz, F.A., 2005. Leaf-level physiology, biomass, and reproduction of *Phytolacca americana* under conditions of elevated CO_2 and altered temperature regimes. Int. J. Plant Sci. 166, 615–622.

Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54, 187–211.

Huxman, T.E., Hamerlynch, E.P., Smith, S.D., 1999. Reproductive allocation and seed production in *Bromus madritensis* ssp. *rubens* at elevated atmospheric CO_2. Funct. Ecol. 13, 769–777.

Machin, D., Sanderson, B., 1977. Computing maximum-likelihood estimates for the parameters of the de Wit competition model. Appl. Stat. 26, 1–8.

Mc Connaughty, K.D.M., Coleman, J.S., 1999. Biomass allocation in plants: ontogeny or optimality? A test along three resource gradients. Ecology 80, 2581–2593.

Mc Lachlann, P., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons Inc., New York.

Mendez, M., Karlsson, P.S., 2004. Between-population variation in size-dependent reproduction and reproductive allocation in *Pinguicula vulgaris* (*Lentibulariaceae*) and its environmental correlates. Oikos 104, 59–70.

Quinn, G.P., Keough, M.J., 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press.

Rajakaruna, N., Bradfield, G.E., Bohm, B.A., Whitton, J., 2003. Adaptive differentiation in response to water stress by edaphic races of *Lasthenia californica* (Asteraceae). Int. J. Plant Sci. 164, 371–376.

Schmid, B., Polasek, W., Weiner, J., Krause, A., Stoll, P., 1994. Modeling of discontinuous relationships in biology with censored regression. Am. Nat. 143, 494–507.

Schneider, M.K., Law, R., Illian, J.B., 2006. Quantification of neighbourhood-dependent plant growth by Bayesian hierarchical modelling. J. Ecol. 94, 310–321.

Schwartz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Sletvold, N., 2002. Effects of plant size on reproductive output and offspring performance in the facultative biennial *Digitalis purpurea*. J. Ecol. 90, 958–966.

Sokal, R.R., Rohlf, F.J., 1995. Biometry. Wiley & Sons, NY.

Stanton, M.L., Roy, B.A., Thiede, D.A., 2000. Evolution in stressful environments. I. Phenotypic variability, phenotypic selection, and response to selection in five distinct environmental stresses. Evolution 54, 93–111.

Sugiyama, S., Bazzaz, F.A., 1998. Size dependence of reproductive allocation: the influence of resource availability, competition and genetic identity. Funct. Ecol. 12, 280–288.

Thompson, B.K., Weiner, J., Warwick, S.I., 1991. Size-dependent reproductive output in agricultural weeds. Can. J. Bot. 69, 442–446.

Underwood, A.J., 1997. Experiments in Ecology. Cambridge University Press.

Vega, C.R.C., Sadras, V.O., Andrade, F.H., Uhart, S.A., 2000. Reproductive allometry in soybean, maize and sunflower. Ann. Bot. 85, 461–468.

Warwick, S.I., Beckie, H.J., Thomas, A.G., McDonald, T., 2000. The biology of Canadian weeds 8. *Sinapis arvensis* L. (updated). Can. J. Plant Sci. 80, 939–961.

Wayne, P.M., Carnelli, A.L., Connolly, J., Bazzaz, F.A., 1999. The density dependence of plant responses to elevated CO_2. J. Ecol. 87, 183–192.

Weiner, J., 2004. Allocation, plasticity and allometry in plants. Perspect. Plant Ecol. 6, 207–215.

Wesselingh, R.A., de Jong, T.J., 1995. Bidirectional selection on threshold size for flowering in *Cynoglossum officinale* (Hounds Tongue). Heredity 74, 415–424.