UNDER REVIEW

RESCIENCE C

Replication / Ecology

[Re] Modeling Insect Phenology Using Ordinal Regression and Continuation Ratio Models

Philipp H. Boersch-Supan[1,2]

[1]British Trust for Ornithology, Thetford, United Kingdom – [2]University of Florida, Gainesville, FL, USA

Edited by
(Editor)

Received
–

Published
–

DOI
–

1 Introduction

Phenology, the timing of seasonal biological phenomena, is a key aspect of plant and animal life. It defines the timing and duration of growth and reproduction and thereby determines the ability to capture seasonally variable resources [1].
Phenological analyses often focus on the timing of particular events, such as the dates of peak plant flowering [2].However, for many biological phenomena exact dates of particular events are more difficult to observe than the state of the system itself. For example, repeated but sparse survey visits may record whether a plant is in bud, flowering, or setting fruit, but not the exact dates when each of those stages was reached. Such observations can be used to categorize an organism's state into discrete classes which usually follow natural ordering, e.g. from least to most developed. The resulting data can be described using ordinal regression models [3, 4].
I here replicate a number of ordinal regression models that were developed by Dennis, Kemp, and Beckwith[5] and Candy[6] to describe insect phenology.

2 Data

The models replicated in this study are fitted to a data set on the phenology of the western spruce budworm *Choristoneura freemani* (Lepidoptera: Tortricidae), a defoliating moth that is widespread in western North America [7]. This data set was originally published in [5] and is a subset of a larger budworm survey data set analysed in [8]. The data consist of 12 sampling occassions at which counts of individual budworms in each of seven development stages (five larval instars, pupae, and adults) were recorded. The only available covariate is a measure of seasonal progression, the accumulated degree days calculated using a threshold of 5.5°C. Candy[6] noted an inconsistency in these data, namely that the reported total number of individuals did not correspond to the sum across the seven development stages for two of the sampling occasions. I therefore use the data set as it was republished in [9], where numbers in each stage have been assumed correct and the totals for each sampling occasion were adjusted accordingly.

Correspondence should be addressed to Philipp H. Boersch-Supan (pboesu@gmail.com)
The authors have declared that no competing interests exist.
Code is available at https://github.com/pboesu/replication_candy_1991..
Data is available at https://github.com/pboesu/replication_candy_1991.

3 Methods

The statistical models replicated here are different types of ordinal regression models [4] all with the aim of predicting the proportion of an insect population in a particular development stage at any given given time. In particular, they represent three different parametrisations of the so-called cumulative model and one version of the so-called sequential model. A recent summary of the theory underlying these models is provided in [10].
The models generally assume that the development of an insect follows an unobservable stochastic process $S(t)$ consisting of accumulated increments of development over time t. As the amount of $S(t)$ increases, the insect passes through successive stages, delimited by moults, with the jth moult occuring when the development threshold a_j is reached:

$$\begin{aligned}
\text{stage } 1: &\quad S(t) \leq a_1 \\
\text{stage } 2: &\quad a_1 < S(t) \leq a_2 \\
&\vdots \\
\text{stage } r-1: &\quad a_{r-2} < S(t) \leq a_{r-1} \\
\text{stage } r: &\quad a_{r-1} < S(t)
\end{aligned}$$

The a_j values are typically unknown and must be estimated from the data.

3.1 Cumulative model with constant variance

If the cumulative number of individuals observed in stages 1 to j is given by $m_{ij} = \sum_{k=1}^{j} n_{ik}$ then the ordinal regression model [3] is specified by

$$\begin{aligned}
\mathbf{E}(m_{ij}) &= N_i Pr(S(t) < \alpha_j), \qquad j = 1, \ldots, r &(1)\\
&= N_i G(\alpha_j + \beta z_i) &(2)
\end{aligned}$$

where G is the cumulative probability density function of $S(t)$, α_j are ordered thresholds or cut-point parameters, β is a vector of regression parameters and z_i is a vector of predictor variables. If the probability of an individual being in stage j or earlier at time t_i is

$$\mu_{ij} = \mathbf{E}(m_{ij})/N_i$$

one can define G^{-1} as the link function of a generalised linear model with the linear predictor

$$\eta_{ij} = \alpha + \beta z_i$$

. This ordinal regression model is commonly known as the cumulative model, and is applied to the budworm data in [6] using the logit and complementary log-log link functions. In both cases the parametrisation results in a constant variance for $S(t)$. Candy[6] reexpresses the model in terms of stage-specific counts n_{ij}

$$\mathbf{E}(n_{ij}) = N_i\{G(\alpha_j + \beta z_i) - G(\alpha_{j-1} + \beta z_i)\} \tag{3}$$

and fits it using a Poisson likelihood [11]. No code is provided for this estimation procedure in the original paper, however, [9] provides a set of example macros for the software package `GLIM` [12] which is no longer actively developed or distributed. I therefore created an R version of the estimation procedure which directly optimizes a Poisson log-likelihood for (3) using the `optim` function. The cumulative model is also implemented in various R packages, including `VGAM` [13] and `ordinal` [14] and I here make use of these to fit the model to the budworm data.

3.2 Cumulative model with proportional variance

Dennis, Kemp, and Beckwith[5] proposed a different parametrisation of the ordinal model is based on assuming a logistic distribution for $S(t)$, such that the probability that an insect's development at time t has not exceeded s amounts to

$$Pr[S(t) \leq s] = 1 \Big/ \left\{1 + \exp\left[-\left(\frac{s-t}{\sqrt{b^2 t}}\right)\right]\right\} \tag{4}$$

where b^2 is a positive constant which also must be estimated from the data. This distribution has a mean of t and a variance of $(\pi^2/3)b^2t$. At any fixed time t the thresholds a_j segment the porbaility distribution function in to r parts and the area under the curve between a_{i-1} and a_i gives the probability that the insect will be in stage i at time t.
This modelling approach is applied to a dataset consisting of samples that record the number of insects x_{ij} in stage j at times $t_1, t_2, \ldots, t_q$ and the x_{ij} are assumed to be random samples from a multinomial distribution with corresponding multinomial probabilities p_{ij}

$$p_{ij} = Pr[a_{j-1} < S(t_i) \leq a_j] \tag{5}$$

$$= 1 \Big/ \left\{1 + \exp\left[-\left(\frac{a_j - t_i}{\sqrt{b^2 t_i}}\right)\right]\right\} - 1 \Big/ \left\{1 + \exp\left[-\left(\frac{a_{j-1} - t_i}{\sqrt{b^2 t_i}}\right)\right]\right\} \tag{6}$$

To fulfill the constraint that $\sum_{j=1}^{r} p_{ij} = 1$ it is further assumed that $a_0 = -\infty$ and $a_r = +\infty$. The model has r unknown parameters $a_1, \ldots, a_{r-1}$ and b^2 which can be found by maximising the corresponding log-likelihood function which takes the form

$$\ell = \log C + \sum_{j=1}^{r} \sum_{i=1}^{q} x_{ij} \log p_{ij} \tag{7}$$

where C is a combinatorial constant that is independent of the parameter values.
Dennis, Kemp, and Beckwith[5] provided SAS code to estimate the parameters under this likelihood using an interatively reweighted non-linear least squares approach based on `PROC NLIN`. This code only required minimal updates to run in a contemporary version of SAS (SAS 9.4) and is provided in the article repository. However, since SAS is a proprietary software package, I created an R version of the estimation procedure which directly optimizes the log-likelihood (7) using the `optim` function and initial values provided in [5].
Candy[6] re-expresses (6) to match the form of (3), which results in the following reparameterisation $\alpha_j = a_j/b$, $\beta = -1/b$, and $z_i = \sqrt{t_i}$, and uses the poisson likelihood approach described above for parameter estimation.

3.3 Sequential model

A different class of ordinal regression models, the sequential model, can be derived by treating the observations as the result of a strictly ordinal counting process, in the sense that to achieve a stage j, all lower stages $1, \ldots, j-1$ have to be achieved. The general form of this model is known as the sequential model, and rather than assuming a single latent process $S(t)$ as in the cumulative model there is a latent continuous variable S_j for each category j. As in the cumulative model this can be framed as a GLM

$$S_j = \eta + \epsilon_j \tag{8}$$

with a linear predictor η and an error term ϵ_j which has mean zero and is distributed following some distribution G. This leads to a model of the form

$$Pr[S = j | S \geq j, \eta] = G(a_j - \eta) \tag{9}$$

Table 1. Parameter estimates for the cumulative logit model with proportional variance. This table replicates results presented in the first row of Table 1 of [8] and the last row of Table 2 of [6].

a_1	a_2	a_3	a_4	a_5	a_6	b^2	Method	Eqn.
121.080	204.360	264.410	342.473	465.620	599.570	1.559	Original [8]	6
120.000	204.700	264.600	341.300	464.500	595.700	1.412	SAS `NLIN`	6
120.033	204.659	264.586	341.285	464.480	595.690	1.412	R `optim`	6
α_1	α_2	α_3	α_4	α_5	α_6	β	Method	Eqn.
101.000	172.200	222.700	287.200	390.900	501.300	-0.842	Original [6]	3
100.990	172.181	222.598	287.134	390.771	501.157	-0.841	R `optim`	3

Table 2. Parameter estimates for the cumulative model with constant variance. This table replicates results presented in the first two rows of Table 2 of [6].

α_1	α_2	α_3	α_4	α_5	α_6	β	Link	Method
5.49	9.39	12.23	15.70	21.26	27.25	-0.05	logit	Original [6]
5.53	9.48	12.35	15.86	21.46	27.52	-0.05	logit	R `optim`
5.47	9.36	12.21	15.67	21.22	27.19	0.05	logit	R `clm`
5.47	9.36	12.21	15.67	21.22	27.19	-0.05	logit	R `vglm`
3.32	5.85	7.72	10.03	13.46	17.52	-0.03	cloglog	Original [6]
3.48	6.10	8.05	10.45	14.01	18.23	-0.03	cloglog	R `optim`
3.32	5.85	7.71	10.02	13.46	17.52	0.03	cloglog	R `clm`
NA	NA	NA	NA	NA	NA	NA	cloglog	R `vglm`

When G is the logistic distribution this model is also known as the continuation ratio model [15]. Confusingly, there are two common versions of the model in the literature both using this name. The one outlined above describing the probability of the sequential process *stopping* at stage j, and the other describing the probability of the process *continuing* beyond stage j, i.e. $Pr(Y > j|Y \geq j)$ [10, 13]. The paper replicated here [6] used the stopping parametrisation of this model. In their notation this leads to an expected value for the stage-specific counts n_{ij}

$$\mathbf{E}(n_{ij}) = N_i G(\beta_{01} + \beta_{11} t_i), \qquad j = 1 \tag{10}$$

$$= N^*_{ij} G(\beta_{0j} + \beta_{1j} t_i), \qquad j = 2, \ldots, r-1 \tag{11}$$

$$\tag{12}$$

with

$$N^*_{ij} = \left(N_i - \sum_{k=1}^{j-1} n_{ik} \right)$$

and conditional probabilities
Candy[6] uses a maximum likelihood estimation to fit these models assuming the n_{ij} are binomially distributed conditional on N_i for stage 1 and conditional on N^*_{ij} for stages $2, \ldots, r-1$. No code is provided for this estimation procedure in the original paper. However, the sequential model with stopping parameterisation is implemented in `VGAM` [13] and I here make use of it to fit the model to the budworm data.

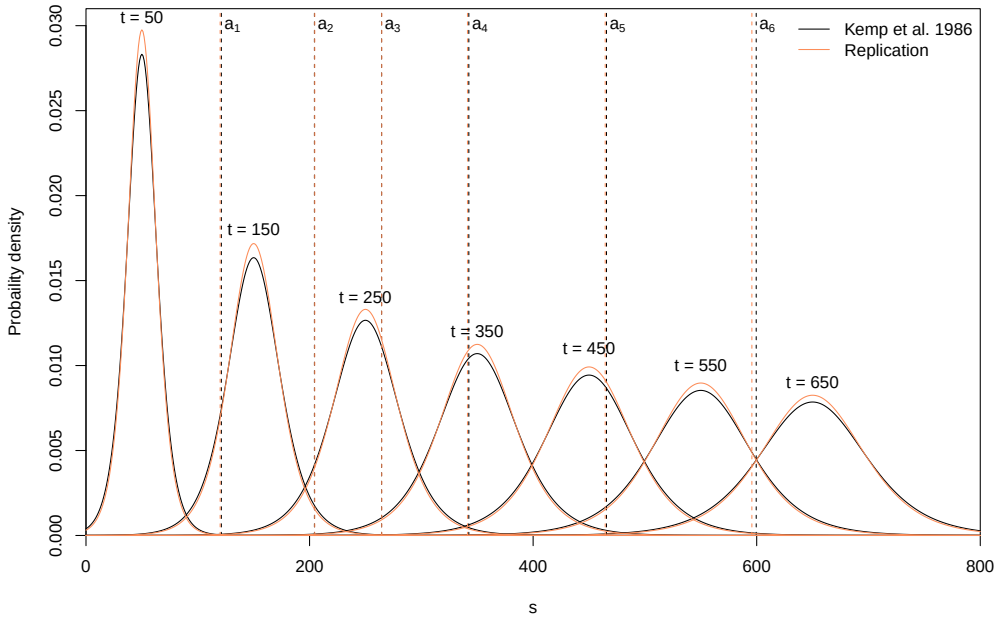

Figure 1. Logistic PDF of the Dennis, Kemp, and Beckwith[5] model plotted for seven fixed values of t. Area under the PDF between a_{j-1} and a_j gives the expected proportion of insects in stage j at time t. Values of a_j and b^2 used in the graph are the estimates given in Table 1 of [8] (black lines) and the estimates from the replication (red lines). This figure replicates Figure 2 in [5].

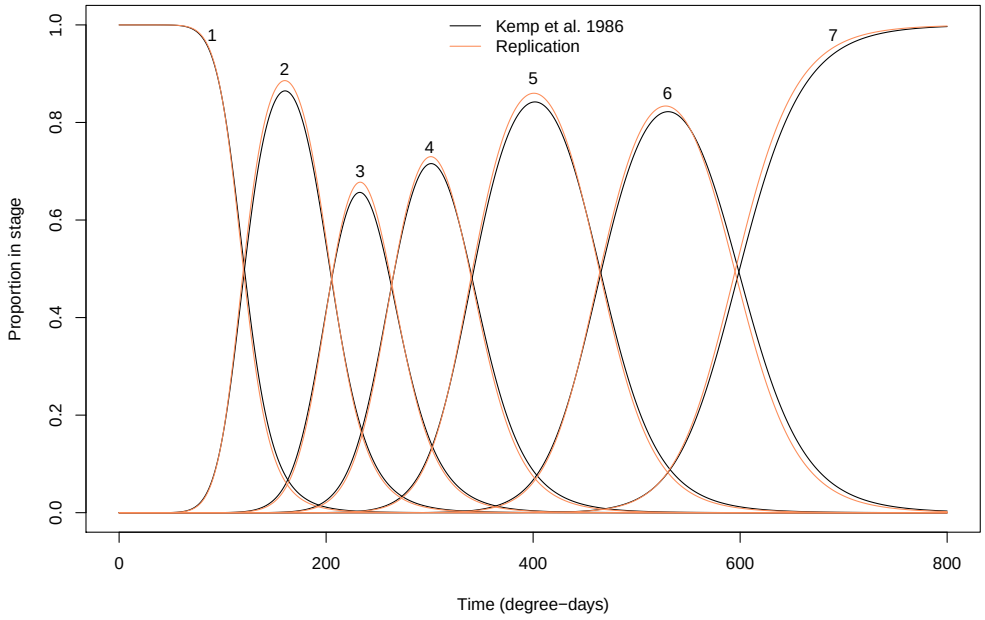

Figure 2. Expected proportion of insects in stages 1-7 plotted as functions of time t. Values of a_j and b^2 used in the graph are the estimates given in Table 1 of [8] (black lines) and the estimates from the replication (red lines). This figure replicates Figure 3 in [5].

4 Results

5 Results

References

1. I. Chuine and J. Régnière. "Process-Based Models of Phenology for Plants and Animals." In: **Annual Review of Ecology, Evolution, and Systematics** 48.1 (2017), pp. 159–182.
2. Y. Aono and K. Kazui. "Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century." In: **International Journal of Climatology: A Journal of the Royal Meteorological Society** 28.7 (2008), pp. 905–914.
3. P. McCullagh. "Regression Models for Ordinal Data." In: **Journal of the Royal Statistical Society: Series B (Methodological)** 42.2 (Jan. 1980), pp. 109–127. DOI: 10.1111/j.2517-6161.1980.tb01109.x. URL: https://doi.org/10.1111%2Fj.2517-6161.1980.tb01109.x.
4. A. Agresti. **Analysis of Ordinal Categorical Data**. John Wiley & Sons, Inc., Mar. 2010. DOI: 10.1002/9780470594001. URL: https://doi.org/10.1002%2F9780470594001.
5. B. Dennis, W. P. Kemp, and R. C. Beckwith. "Stochastic Model of Insect Phenology: Estimation and Testing." In: **Environmental Entomology** 15.3 (June 1986), pp. 540–546. DOI: 10.1093/ee/15.3.540. URL: https://doi.org/10.1093%2Fee%2F15.3.540.
6. S. G. Candy. "Modeling insect phenology using ordinal regression and continuation ratio models." In: **Environmental entomology** 20.1 (1991), pp. 190–195.
7. M. H. Brookes, R. W. Campbell, J. J. Colbert, R. G. Mitchell, and R. W. Stark. **Western spruce budworm**. Cooperative State Research Service Technical Bulletin 1694. United States Department of Agriculture Forest Service, 1987.
8. W. P. Kemp, B. Dennis, and R. C. Beckwith. "Stochastic Phenology Model for the Western Spruce Budworm (Lepidoptera: Tortricidae)." In: **Environmental Entomology** 15.3 (June 1986), pp. 547–554. DOI: 10.1093/ee/15.3.547. URL: https://doi.org/10.1093%2Fee%2F15.3.547.
9. S. G. Candy. "Biology of the mountain pinhole borer, Platypus subgranosus Schedl, in Tasmania." PhD thesis. University of Tasmania, 1990.
10. P.-C. Bürkner and M. Vuorre. "Ordinal Regression Models in Psychology: A Tutorial." In: **Advances in Methods and Practices in Psychological Science** 2.1 (Feb. 2019), pp. 77–101. DOI: 10.1177/2515245918823199. URL: https://doi.org/10.1177%2F2515245918823199.
11. R. Thompson and R. Baker. "Composite link functions in generalized linear models." In: **Journal of the Royal Statistical Society: Series C (Applied Statistics)** 30.2 (1981), pp. 125–131.
12. M. Aitkin. **Statistical modelling in GLIM**. Oxford Oxfordshire New York: Clarendon Press Oxford University Press, 1989.
13. T. W. Yee. "The VGAM Package for Categorical Data Analysis." In: **Journal of Statistical Software** 32.10 (2010), pp. 1–34. URL: http://www.jstatsoft.org/v32/i10/.
14. R. H. B. Christensen. **ordinal—Regression Models for Ordinal Data**. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal. 2019.
15. S. E. Fienberg. "The analysis of cross-classified categorical data." In: **Massachusetts Institute of Technology Press, Cambridge and London** (1980).