COMPUTER SCIENCE AND STATISTICS:
The Interface, D.M. Allen (ed.)

GRAPHICAL ANALYSIS OF PROPORTIONAL POISSON RATES

Brian S. Yandell

University of Wisconsin - Madison

We present graphical tools for examining proportionality of a Poisson process rate to a baseline from a group of similar processes. We examine smooth deviations from this baseline using smoothing splines for general linear models. An example of egg-laying rates for leafhoppers is examined in some detail.

1. Introduction

This paper concerns inference for nonstationary Poisson rates which are "almost" proportional to a common baseline. It provides a means for "pre-smoothing" rate estimates to avoid some of the common problems of estimating functions with large curvature at certain places.

One may believe that a group of female potato leafhoppers in the same fluctuating temperature regime (Hogg, 1984) would oviposit at rates which rose and fell at roughly the same time. That is, one would suppose that the oviposition rates would be proportional to a common baseline rate. One could estimate this baseline rate, and then estimate the individual curves by simply determining the constant of proportionality, as was done by Bartoszyński et al. (1981). However, one might want to examine the proportionality as a function of time to determine whether or not it is constant.

We propose a method to estimate this proportionality over time. Although many approaches are possible (Clevenson and Zidek, 1977; Hastie and Tibshirani, 1984), we develop our estimators in the framework of penalized maximum likelihood (Good and Gaskins, 1971; O'Sullivan, Yandell, and Raynor, Jr., 1984).

Section 2 formulates the problem of proportional rates. Penalized maximum likelihood estimators for the baseline rate and proportionality terms are developed in Section 3. Section 4 briefly presents diagnostic tools. The methods are applied to leafhopper oviposition data in Section 5.

2. Proportional Poisson Rates

An individual leafhopper i, $i=1, \cdots, r$, may lay Y_{ij} eggs at time t_j, $t_1 < \cdots < t_n$. The count Y_{ij} is assumed Poisson with mean $h_i(t_j)$, which may be nonstationary. We focus on the model

$$h_i(t) = h^o(t)\, a_i(t),\ t \geq 0,\ i=1, \cdots, r. \qquad (2.1)$$

Proportional rates would correspond to constant a_i, with $h^o(\cdot)$ being the baseline rate. Taking logarithms yields

$$\log(h_i(t)) = \log(h^o(t)) + \log(a_i(t)), \qquad (2.1)$$

or, reparameterizing one has

$$\theta_i(t)) = \theta^o(t) + \alpha_i(t),\ t \geq 0,\ i=1, \cdots, r. \qquad (2.2)$$

The degree to which the a_i, or α_i, are not constant corresponds to how much the proportional rates assumption is violated. This suggests that one could evaluate the degree of nonproportionality by estimating a_i, or equivalently α_i, and plotting these against time.

2.1. Log Likelihood

The likelihood can be written down and decomposed into pieces so that, subject to constraints, we can have a separate likelihood for each individual proportionality term. The overall log likelihood

$$\frac{1}{nr} \sum_i \sum_j Y_{ij} [\log(Y_{ij}) - \theta^o(t_j) - \alpha_i(t_j)]$$

can be reexpressed as the sum of

$$L(\theta^o) = \frac{1}{n} \sum_j Y_{+j} [\log(Y_{+j}/r) - \theta^o(t_j)] \qquad (2.3)$$

and

$$\frac{1}{nr} \sum_i \sum_j Y_{ij} [\log(r Y_{ij}/Y_{+j}) - \alpha_i(t_j)] \qquad (2.4)$$

Throughout this paper, "+" indicates sum over the intended index. Note that (2.3) is a Poisson penalized likelihood, and (2.4) is a multinomial penalized likelihood conditional on Y_{+j}. In other words, Y_{ij} is binomial $(Y_{+j}, a_i(t_j)/r)$. This suggests splitting (2.4) into r terms of the form

$$L(\alpha_i) = \frac{1}{n} \sum_j Y_{ij} \log\left(\frac{Y_{ij}}{\mu_{ij}}\right) + (Y_{+j} - Y_{ij}) \log\left(\frac{Y_{+j} - Y_{ij}}{Y_{+j} - \mu_{ij}}\right)$$

in which $\mu_{ij} = a_i(t_j) Y_{+j}/r$. Thus the log likelihood can be split into $r+1$ terms, for θ^o and for α_i, $i=1, \cdots, r$, with the the restriction that $\sum a_i(t)/r = 1$.

3. Penalized Maximum Likelihood Estimates

We now impose a penalty on the estimators to insure a certain smoothness not guaranteed by the likelihood as written. The penalized maximum likelihood estimate (MPLE) for the baseline rate (Bartoszyński et al., 1981; O'Sullivan, Yandell, and Raynor, Jr., 1984) can be found by minimizing, for fixed λ,

$$L(\theta^o, \lambda) = L(\theta^o) + \lambda J(\theta^o) \qquad (3.1)$$

in which J() is an appropriate penalty function, typically

$$J(f) = \int (f^{(m)}(t))^2\, dt \qquad (3.2)$$

with m = 1 or 2 for penalty on the slope or curvature, respectively. A large value of the penalty, or smoothing, parameter λ forces θ^o to be nearly linear, while a small λ allows θ^o to interpolate the data.

The smoothing splines incorporate a prior belief that the true curve is smooth in a certain sense. The smoothing parameters λ are chosen by means of generalized cross validation (Craven and Wahba, 1979), which tries to minimize the mean

square error, forcing a tradeoff between bias and variance.

Similar expressions can be written down for determining the MPLE of α_i for each i,

$$L(\alpha_i,\lambda_i) = L(\alpha_i) + \lambda_i J(\alpha_i) \tag{3.3}$$

When $m=1$ and $\lambda_i=\infty$, the constant MPLEs are

$$\bar{\alpha}_i = \log(rY_{i+}/Y_{++}),\ i=1,\cdots,r.$$

The estimation problem can be split into $r+1$ minimization problems, for θ'' and for α_i, $i=1,\cdots,r$, provided we are willing to ignore the restriction that $\sum a_i/r=1$. Of course, such a restriction could be imposed, but it would place awkward constraints on the smoothing penalties.

3.1. Data over Time Intervals

The data considered in Section 5 is grouped by 2 or 3 day intervals. With this design imbalance, the estimates of θ'' and of α_i may be biased, depending on the pattern of grouping. However, the unconditional expectation of the estimates is unbiased provided that the pattern of grouping is independent of the state of an individual. We can adjust the penalized likelihood expressions in a natural way to account for the reduced data, namely,

$$L(\theta'') = \frac{1}{n}\sum_j Y_{+j}[\log(Y_{+j}/d_{+j}) - \theta''(t_j)] \tag{3.4}$$

$$L(\alpha_i) = \frac{1}{n_i}\sum_j Y_{ij}\log\left(\frac{Y_{ij}}{\mu_{ij}}\right) + (Y_{+j}-Y_{ij})\log\left(\frac{Y_{+j}-Y_{ij}}{Y_{+j}-\mu_{ij}}\right)$$

in which $\mu_{ij} = a_i(t_j)d_{ij}Y_{+j}/d_{+j}$. That is, for each i, there were n_i distinct times t_j at which counts Y_{ij} were made. These counts encompass d_{ij} days each, and the proportion of days for i out of the total count Y_{+j} is d_{ij}/d_{+j}. These technical adjustments were used for computing, but are not pursued further in this paper.

3.2. Survival and Oviposition

Throughout the leafhopper study, individuals died. Thus group size declined over time. These deaths can affect the estimate of the "baseline" rate h^0, as well as the proportionality terms a_i, even if all the rates are constant. This problem is most profound for small groups, such as in the latter portion of the leafhopper experiment.

A simple solution shown in the data analysis section is to factor out a step function from the baseline rate, with steps at times of death. This can be easily accomplished with partial splines (Shiau, 1985; Wahba, 1983a). Appropriate modifications can then be made to (2.4) based on the estimated step sizes. A serious danger arises in overparameterizing the model with steps for each individual.

4. Diagnostics for Poisson Rates

We propose an ad hoc "confidence interval" and log likelihood residuals for graphical inspection of proportionality. At present we have no concrete results, but support these tools by analogy to other work.

Several diagnostics have been proposed for penalized maximum likelihood in the linear (least squares) model with i.i.d. errors. Wahba (1983b) proposed pointwise confidence intervals based on a Bayesian model with normal errors. Carmody, Eubank, and Thombs (1984) proposed jackknife confidence intervals which performed poorly in comparison to the intervals of Wahba (1983b). Other diagnostics based on residuals (Eubank, 1984; Gunst and Eubank, 1983) naturally extend diagnostics for unpenalized problems. Recent work of Cox (1984) offers strong approximation of the penalized least squares estimator in the i.i.d. case, under certain conditions on the design points and smoothing parameter, which lead to simultaneous confidence bands if one ignores bias. Another direction based on a supremum penalty for the regression function (Knafl, Sacks, and Ylvisaker, 1983ab) yields bias-corrected simultaneous confidence bands: here, bias is accounted for by a bias correction.

We adapt Wahba (1983b) to the non-i.i.d. case and argue in an ad hoc fashion that this might have reasonable properties for our problem. We consider the model

$$X = g + \epsilon,\ g \sim N(0,(n\lambda)^{-1}\Sigma_{gg}),\ \epsilon \sim N(0,\Sigma),$$

with Σ diagonal. The posterior estimator of g is

$$\hat{g} = E(g \mid X) = \Sigma_{gg}(\Sigma_{gg}+n\lambda\Sigma)^{-1}X = H_\lambda X. \tag{4.1}$$

The covariance is derived in an analogous fashion as

$$COV(g \mid X) = (I+H_\lambda)\Sigma_{gg}/(n\lambda) = H_\lambda\Sigma. \tag{4.2}$$

This suggests an approximate 95% confidence interval for g_j

$$\hat{g}_j \pm 1.96\,\sigma_j\sqrt{h_{jj}(\lambda)} \tag{4.3}$$

Now suppose, for fixed i, we let $X_j=\log(Y_{ij}/(Y_{+j}-Y_{ij}))$ and approximate the covariance to first order,

$$\sigma_j^2 = 2r\exp(-\alpha_i(t_j))/Y_{+j},\ j=1,\cdots,n.$$

The estimated confidence interval for $\alpha_i(t_j)$ becomes

$$\hat{\alpha}_{i\lambda}(t_j) \pm 1.96\sqrt{2h_{jj}(\lambda_i)r\exp(-\hat{\alpha}_{i\lambda}(t_j))/Y_{+j}} \tag{4.4}$$

This approach has some problems, as the solution to the penalized log likelihood is not the same as the solution to a logit regression with normal errors. We will pursue this in later work using ideas of Leonard (1982).

We propose an ad-hoc test of the hypothesis of constant proportionality by computing the difference in deviances between the smooth and constant estimates,

$$D(i,\lambda) = 2[L(\bar{\alpha}_i) - L(\hat{\alpha}_{i\lambda})],\ i=1,\cdots,r, \tag{4.5}$$

with $\hat{\alpha}_{i\lambda}()$ being the spline estimate of $\alpha_i()$ for fixed smoothing parameter λ and $\bar{\alpha}_i$ the estimate for constant α_i. In other words, $D(i,\lambda)$ is simply the deviance between the constant and the smoothed logit models. We suppose that this statistic may have approximately a chi-square distribution with degrees of freedom $(n-1)-trace(I-H_\lambda)$. We will compare this with the usual likelihood ratio statistic, $D(i)=2L(\bar{\alpha}_i)$ with $n-1$ degrees of freedom, in the data analysis section.

Expression (4.5) suggests examining the deviance contributions at t_j (Green, 1984; Pregibon, 1981)

$$\pm[2Y_{ij}(\log(Y_{ij}) - \hat{\alpha}_{i\lambda}(t_j))]^{1/2}, \tag{4.6}$$

with the sign the same as that of $Y_{ij}-\exp(\hat{\alpha}_{i\lambda}(t_j))Y_{+j}/r$. For given t_j, this is approximately $N(0,1)$; thus large positive or negative values suggest significant deviations. However, the graphical "tests" at different t_j are highly correlated, and a graphical plot of t versus logit residuals cannot be viewed as a global test.

5. Data Analysis

We consider data from a laboratory experiment conducted by Hogg (1984) in which female potato leafhoppers were kept in controlled laboratory conditions at one of three fluctuating tem-

perature regimes. We focus here only on the cold regime. We examine the baseline for the 23 females in this group along with the proportional term for two of these females. A more complete analysis is in progress jointly with David Hogg, Entomology Department, UW-Madison, who kindly offered the data he collected.

All individuals have grouped records, that is counts of eggs for 1-3 day intervals. Also, individuals were removed from the study by death, either natural or accidental (due to handling). We assume that the grouping does not introduce any bias in the estimation of the baseline rate, and that we are interested in the baseline rate and proportionality terms at any time only for those leafhoppers which were alive. We initially proceed as if survival did not affect bias, and later correct for survival as indicated in Section 3.3.

Figure 5.1 shows the baseline rate and the rates for individuals 22 and 23. Note the rise to a fairly constant rate, with gradual decay. The raw proportionality for individuals 22 and 23 are plotted alongside curve estimates with penalties for slope and for curvature in Figures 5.2-3. The curve estimate based on a penalty for non-zero slope appear much rougher than the curves based on curvature penalty. Approximate 95% pointwise confidence intervals for the proportionality estimates, based on the curvature penalty, are shown in Figures 5.4-5.

The likelihood ratio statistics with degrees of freedom and p-value are shown in Table 5.1. Note the great reduction in degrees of freedom for the penalized curves, while the deviances stay fairly high. Figure 5.6-7 show the logit deviances over time.

Table 5.1 Smooth Deviances

	Deviance	d.f.	log(λ)
#22:			
constant	188.21	68.	∞
m = 1 (slope)	117.66	14.88	-6
m = 2 (curvature)	99.49	8.83	-12
#23:			
constant	113.99	64.	∞
m = 1 (slope)	63.56	5.87	-4
m = 2 (curvature)	62.36	3.35	-8

We conclude with curve estimates for the baseline once one adjusts for the survival process. Figure 5.8 shows the naive and adjusted baseline rate estimates for the cold regime. One sees that survival has little effect on the baseline rate for most of the experiment, though estimates at the later times can be affected.

Acknowledgements

This work was supported in part by USDA-CSRS grant 511-100. Computing was performed on the Statistics VAX 11/750 Research Computer at the University of Wisconsin. Discussions with David Hogg were most helpful.

References

Bartoszyński, R., Brown, B. W., McBride, C., and Thompson, J. R. (1981), "Some Nonparametric Techniques for Estimating the Intensity Function of a Cancer Related Nonstationary Poisson Process," *Ann. Statist.*, 9, 1050-1060.

Carmody, T. J., Eubank, R. L., and Thombs, L. A. (1984) "Jackknife Confidence Intervals for Smoothing Splines." Technical Report, Dept. of Statistics, So. Methodist U..

Clevenson, M. L., and Zidek, J. V. (1977), "Bayes Linear Estimators of the Intensity Function of the Nonstationary Poisson Process," *Journal of the American Statistical Association*, 72, 112-120.

Cox, D. D. (1984) "Gaussian Approximation of Smoothing Splines." Technical Report#743, Dept. of Statistics, U. of Wisconsin.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numer. Math.*, 31, 377-403.

Eubank, R. L. (1984), "The Hat Matrix for Smoothing Splines," *Statist. & Probab. Letters*, 2, 9-14.

Good, I. J., and Gaskins, R. A. (1971), "Non-Parametric Roughness Penalties for Probality Densities," *Biometrika*, 58, 255-277.

Green, P. J. (1984), "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives (with Discussion)," *J. Roy. Statist. Soc. B*, 46, 149-192.

Gunst, R. F., and Eubank, R. L. (1983) "Regression Diagnostics and Approximate Inference Procedures for Penalized Least Squares Estimators." Technical Report#181, Dept. of Statistics, So. Methodist U..

Hastie, T. J., and Tibshirani, R. J. (1984) "Generalized Additive Models." Technical Report#98, Div. of Biostatistics, Stanford U..

Hogg, D. B. (1984) "Potato Leafhopper (Homoptera: Cicadellidae) Immature Development, Life Tables, and Population Dynamics Under Fluctuating Temperature Regimes." *Environmental Entomology*, . (submitted)

Knafl, G., Sacks, J., and Ylvisaker, D. (1983a) "Uniform Confidence Intervals for Regression Estimates." Technical Report, Center for Statistics and Probability, Northwestern U..

Knafl, G., Sacks, J., and Ylvisaker, D. (1983b) "Model Robust Confidence Intervals II." Technical Report#55, Center for Statistics and Probability, Northwestern U..

Leonard, T. (1982) "An Empirical Bayesian Approach to the Smooth Estimation of Unknown Functions." Technical Report#2339, Math. Research Center, U. Wisconsin.

O'Sullivan, F., Yandell, B. S., and Raynor, Jr., W. J. (1984) "Automatic Smoothing of Regression Functions in Generalized Linear Models." Technical Report#734, Dept. of Statistics, U. of Wisconsin.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *Annals of Statistics*, 9, 705-724.

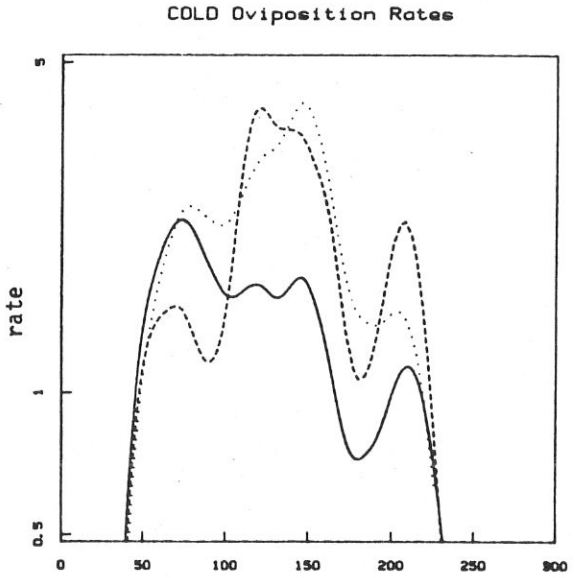
Shiau, J-J H. (1985), "Smoothing Spline Estimation of Functions with Discontinuous D-th Derivatives". (Ph.D. thesis in preparation, U. Wisconsin, Madison)

Wahba, Grace (1983a) "Cross Validated Spline Methods for the Estimation of Multivariate Functions from Data on Func-

tionals." Technical Report#722, Dept. of Statistics, U. of Wisconsin.

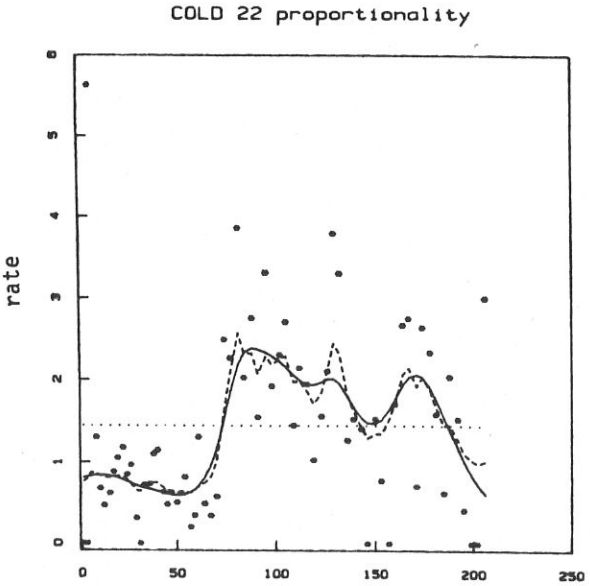
Wahba, G. (1983b), "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *J. Roy. Statist. Soc. B*, 45, 133-150.

Figure 5.1



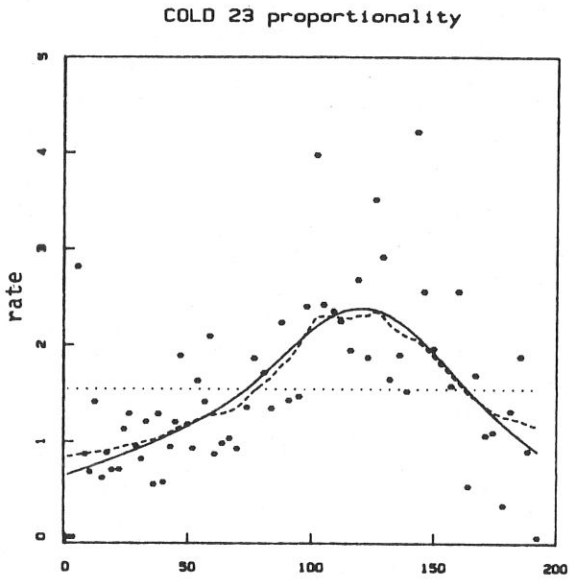
solid=baseline; dash=22; dot=23

Figure 5.2



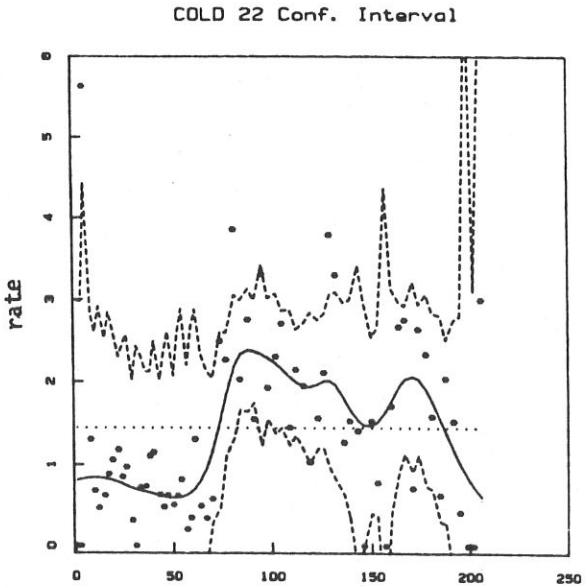
solid=curve; dash=slope; dot=constant

Figure 5.3



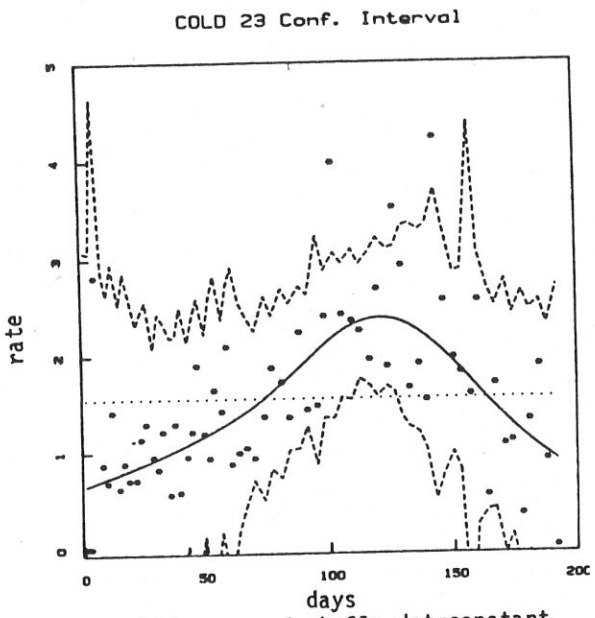
solid=curve; dash=slope; dot=constant

Figure 5.4



solid=curve; dash=CI; dot=constant

Figure 5.5



solid=curve; dash=CI; dot=constant

Figure 5.6

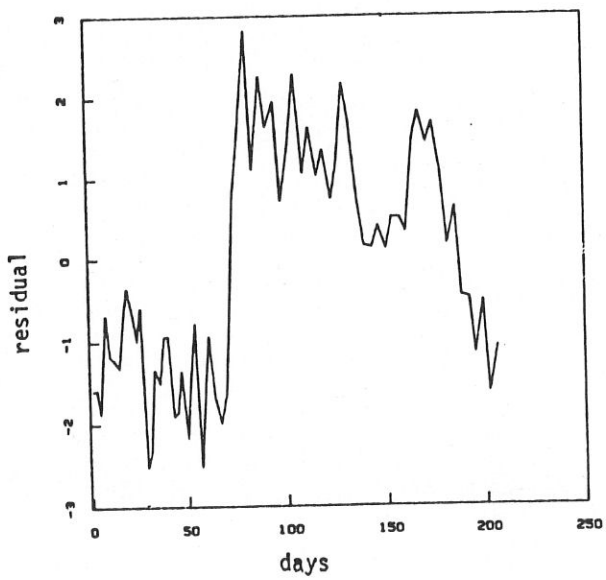


Figure 5.7

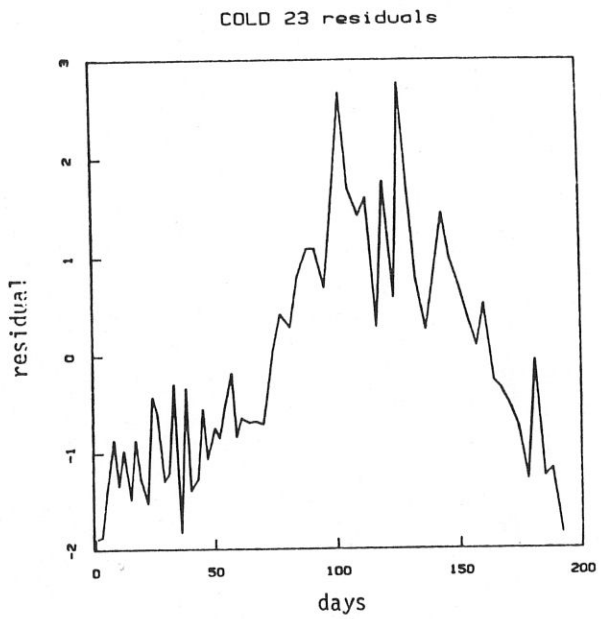
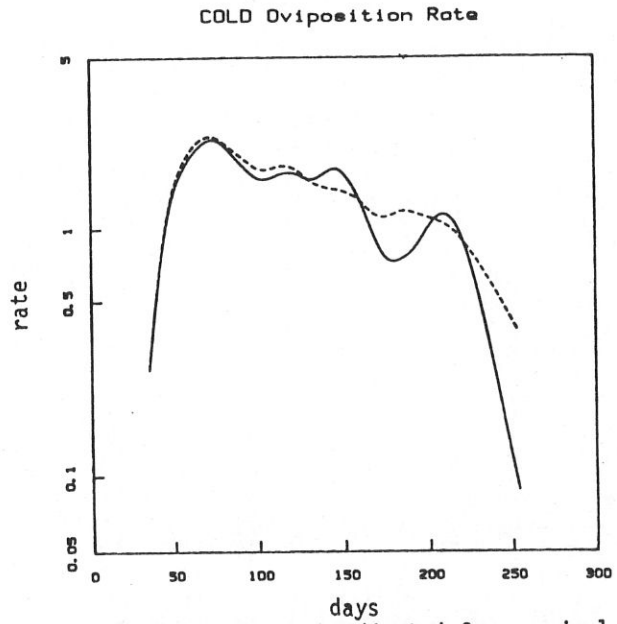


Figure 5.8



solid=smooth; dash=adjusted for survival