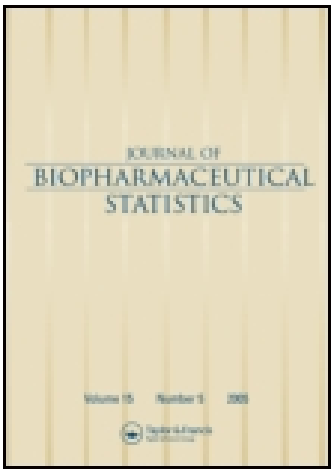

Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

Poisson regression analysis in clinical research

Farid Kianifard [a] & Paul P. Gallo [b]

[a] Biostatistics Department , Hoechst-Roussel Pharmaceuticals , P.O. Box 2500, Somerville , New Jersey , 08876-1258

[b] Biostatistics Department , Ciba Phamaccuticals Division , Summit , New Jersey , 07901

Published online: 29 Mar 2007.

To cite this article: Farid Kianifard & Paul P. Gallo (1995) Poisson regression analysis in clinical research, Journal of Biopharmaceutical Statistics, 5:1, 115-129, DOI: 10.1080/10543409508835101

To link to this article: http://dx.doi.org/10.1080/10543409508835101

Journal of Biopharmaceutical Statistics, 5(1), 115–129 (1995)

POISSON REGRESSION ANALYSIS IN CLINICAL RESEARCH

Farid Kianifard[1] *and Paul P. Gallo*[2]

[1]Biostatistics Department
Hoechst-Roussel Pharmaceuticals
P.O. Box 2500
Somerville, New Jersey 08876-1258

[2]Biostatistics Department
Ciba Pharmaceuticals Division
Summit, New Jersey 07901

Key words. Deviance; Generalized linear model; Overdispersion; Residuals; Sample size

Abstract

Generalized linear models (GLM) are now widely used in analyzing data from clinical trials and in epidemiological studies. In Poisson regression, which fits in the framework of a GLM, the response variable is a count that follows the Poisson distribution. This article describes the basic methodology of Poisson regression analysis and its application to clinical research. Overdispersion, model diagnostics, and sample size issues are discussed. The methodology is illustrated on a data set from a clinical trial for the treatment of bladder cancer, using a new procedure (PROC GENMOD) in the statistical package SAS.

1. Introduction

Generalized linear models (GLM) are now widely used in analyzing data from clinical trials and in epidemiological studies. Perhaps the most familiar in this class, excluding classical linear regression, is the logistic regression model in which the response variable is binary. This article deals with Poisson regression, which also fits in the framework of a GLM. As the name suggests, the response variable is a count that follows the Poisson distribution. Appearing first in 1983, the book by McCullagh and Nelder (1) remains the authoritative text on GLMs and, as such, covers Poisson regression as a special case. A more introductory treatment of GLMs is provided by Dobson (2). Overviews of GLMs and related topics appear in several chapters in a volume edited by Hinkley *et al.* (3). Accounts of GLMs, tied to using statistical packages SAS and GLIM, are given in a report issued by the SAS Institute, Inc. (4) and by Aitkin *et al.* (5), respectively. Although coverage of logistic regression is now standard fare in many texts on linear regression analysis, the same cannot be said about Poisson regression. Two notable exceptions are the texts by Kleinbaum *et al.* (6) and Myers (7), both of which discuss and illustrate the application of Poisson regression with real data.

Much of the work done specifically on Poisson regression centers on papers written by E. L. Frome and his colleagues, the earliest of which was by Frome, Kutner, and Beauchamp (8). Two later papers, by Frome (9) and Frome and Checkoway (10), were instrumental in greater application of Poisson regression in epidemiology. An expository paper was also written for statisticians at large (11).

Several recently published papers demonstrate applications of Poisson regression in medical research. Gallagher *et al.* (12) studied the effect of a therapeutic agent (versus placebo) on patients with postmenopausal osteoporosis, in a clinical trial in which bone fracture frequency was the efficacy variable. Parker (13) presented the results of an analysis of surveillance data on the number of abortion-associated deaths in the United States for the period 1962–1984. Vonesh (14) examined potential risk factors relating to the number of bacterial infections of the peritoneum, to determine whether future randomized clinical trials should have patients stratified according to the identified risk factors before randomization. Campbell *et al.* (15) modeled the factors influencing the number of involuntary expulsions of a vaginal ring used as a method of contraception. Similarly, Poisson regression analysis was used by Rochon *et al.* (16) to examine the number of rejection episodes experienced by renal transplant patients during the first 3 months after transplantation. Neugebauer *et al.* (17) considered the impact of various stressful life events on the seizure frequency of patients with epilepsy.

This article describes the basic methodology of Poisson regression analysis and its application to clinical research. Section 2 defines the model and examines its estimation. Model diagnostics are described in Section 3, along with a discussion of a major concern in the application of Poisson regression, namely, overdispersion. A method for sample size determination in Poisson regression is presented in Section 4. Data from a clinical trial for the treatment of bladder cancer are used in Section 5 to illustrate much of the methodology presented in earlier sections. In addition, the analysis presented illustrates the use of PROC GENMOD, a new procedure in SAS for analysis of GLMs.

2. Model and Estimation

For a Poisson distributed random variable Y with mean μ, the probability of y "events" in unit time is given by

$$P_Y(y; \mu) = (e^{-\mu}\mu^y)/y!, \qquad y = 0, 1, 2, \ldots$$

The model in Poisson regression for each observation y_i ($i = 1, \ldots, n$) can be written as

$$E(y_i) = \mu_i,$$

where μ_i, the mean number of events in a time period of length t_i, is modeled as a function of p explanatory variables (including an intercept) $x_i^T = (x_{i1}, \ldots, x_{ip})$:

$$\mu_i = t_i\lambda(x_i, \beta), \qquad i = 1, \ldots, n. \tag{1}$$

$\beta^T = (\beta_1, \ldots, \beta_p)$ is a vector of unknown parameters and $\lambda(x_i, \beta)$ (>0) is the rate function. Let g be a (differentiable) function of $\lambda(x_i, \beta)$ such that $g(\lambda(x_i, \beta)) = x_i^T\beta$; in the terminology of GLM, g is called the *link function*. Examples of choices for the link function include *log link*, for which $\lambda(x_i, \beta) = \exp(x_i^T\beta)$, and *identity link*, for which $\lambda(x_i, \beta) = x_i^T\beta$. A log link function is the conventional link function in Poisson regression. Since the Poisson distribution values are nonnegative, using a link function whose inverse takes on only nonnegative values is desirable. In addition, as described by McCullagh and Nelder (1, Section 2.2), certain exponential family distributions have corresponding link functions, called *canonical links*, with desirable statistical properties, particularly in small samples; log is the canonical link function for Poisson models. Using the log link, we obtain

$$\ln \mu_i = \ln t_i + x_i^T\beta, \qquad i = 1, \ldots, n. \tag{2}$$

In the above equation, $\ln t_i$ has a coefficient of unity and hence can be absorbed into the left-hand side by regarding $\ln(\mu_i/t_i)$ as the response variable.

More generally, one can include $\ln t_i$ in the model as an additional explanatory variable with an associated parameter β_{p+1}:

$$\ln \mu_i = x_i^T \beta + \beta_{p+1} \ln t_i, \qquad i = 1, \ldots, n. \tag{3}$$

Maximum likelihood estimates, $\hat{\beta}$, of the parameters β can be found by the Newton-Raphson method or an iteratively reweighted least squares procedure, used by SAS and GLIM, respectively.

Let $L(y; \hat{\mu})$ and $L(y; \hat{\beta})$ denote the values of the log likelihood function evaluated at $\mu_i = y_i$ and $\mu_i = t_i \lambda(x_i, \hat{\beta}) = \hat{y}_i$, $i = 1, \ldots, n$, respectively. Then the following quantities measuring the adequacy of the model fit can be defined:

$$\begin{aligned} D(\hat{\beta}) &= 2\{L(y; \hat{\mu}) - L(y; \hat{\beta})\} \\ &= 2\{\Sigma[y_i \ln (y_i/\hat{y}_i) - (y_i - \hat{y}_i)]\}, \qquad \text{(log link)} \end{aligned} \tag{4}$$

$$X^2 = \Sigma[(y_i - \hat{y}_i)^2/\hat{y}_i]. \tag{5}$$

$D(\hat{\beta})$ is called the *deviance* and provides a measure of residual variation for the goodness-of-fit of the assumed Poisson regression model in (1). X^2 is the Pearson's goodness-of-fit statistic. Both quantities are asymptotically distributed as χ^2 statistics, with $(n - p)$ degrees of freedom (d.f.) for model (2) and $(n - p - 1)$ d.f. for model (3).

The deviance has two appealing properties that are not shared by X^2. One is a monotonic decrease in deviance as the number of explanatory variables in the model increases. The other is that, for a hierarchical class of models, the deviance can be decomposed into the sequential contribution of each explanatory variable added to the model.

For hypothesis testing involving parameters of the model, likelihood ratio tests comparing two models in a hierarchical class can be constructed using the deviance. Let $\beta_r^T = (\beta_1, \ldots, \beta_r)$, where $r < p$. Under the null hypothesis H_0: $\beta_{r+1} = \ldots = \beta_p = 0$, the deviance difference $[D(\hat{\beta}_r) - D(\hat{\beta})]$ has an asymptotic χ^2 distribution with $(p - r)$ d.f.

3. Overdispersion and Model Diagnostics

Overdispersion refers to a situation in which the observed variance of a random variable Y exceeds its nominal variance under the assumed probability model. This is often found to occur with Poisson distributed data. If the Poisson regression model is adequate, the expected value of both the deviance and of X^2 is their associated d.f. (n − p), since both have an asymptotic χ^2 distribution. When the deviance or X^2 is noticeably larger than the degrees of freedom, overdispersion is indicated, and doubt is cast on the adequacy of the model.

Some of the possible causes of overdispersion can be examined through residual diagnostics (18), using similar logic as in standard linear models; these include the presence of outliers and model misspecification. Two types of residuals that are useful for Poisson regression are the *deviance residual* (r_{Di}) and the *Pearson residual* (r_{Pi}), defined for $i = 1, \ldots, n$ as

$$r_{Di} = \pm \{2[y_i \ln (y_i/\hat{y}_i) - (y_i - \hat{y}_i)]\}^{1/2}$$

and

$$r_{Pi} = (y_i - \hat{y}_i)/\hat{y}_i^{1/2}.$$

r_{Di} takes the same sign as $(y_i - \hat{y}_i)$. Note that the residual for each observation is simply the square root of the contribution of that observation to the corresponding statistic in Eq. (4) or (5).

Standardized versions of the deviance residuals and Pearson residuals with unit asymptotic variance have also been proposed by Williams (19) and others: the ith residual is divided by $(1 - h_i)^{1/2}$, where h_i is a measure of leverage of the ith observation, obtained as the ith diagonal element of a projection matrix (analogous to the "hat matrix" of linear models; however, in Poisson regression, h_i is a function of the response variable as well as of the explanatory variables). The h_i are likely to be close to zero in the rather large sample sizes often employed in clinical trials, and hence their impact on the size of the deviance residuals and Pearson residuals is minimal. Such standardized residuals are not produced by current versions of available software, such as GLIM or PROC GENMOD in SAS.

Residuals can be used in ways that parallel linear models diagnostics. For example, examination of the above residuals for large values (say, >3 in magnitude) may point to the presence of outliers. It should be noted, however, that there are no corresponding distributional properties for these residuals; i.e., they cannot be assumed to be normally distributed [see, however, Pierce and Schafer (20), Williams (19), and Davison and Gigli (21)]. As in other types of models, outliers should be scrutinized for possible reasons for their aberrance and a justification should be provided if they are to be excluded from the analysis. Adequacy of the model should be questioned when too many observations are judged to be outliers; an acceptable number depends on the particular application, but in general, it could be considered a cause for concern if at least 10–15% of the observations were judged to be outliers.

Residuals can also be used to examine whether the functional form of the model is correct. For example, if the model contains a linear term for a particular continuous explanatory variable, then plotting residuals versus that variable and looking for a systematic pattern can address the appropriateness

of the assumed linear form. Other residual plots with analogs to standard linear models have been proposed by Wang (22,23) and Lee (24).

Absence of important explanatory terms from the model, such as interactions involving terms already in the model, can also result in overdispersion. This can be addressed using the deviance to compare models in a hierarchical class, as described in Section 2.

Overdispersion may result if many of the predicted values of the response variable ($\hat{y}_i$) are small, resulting in a poor asymptotic χ^2 approximation to the distribution of the deviance or X^2. Campbell *et al.* (15) describe a method of grouping the fitted values before computing X^2 in order to improve the χ^2 approximation.

In some sense, it could be argued that in Poisson regression analysis, some degree of overdispersion might always be expected, since in the model as stated above, intersubject variation is not explicitly incorporated and each individual Poisson mean is assumed to be exactly determined by the factors in the model; this argument is closely related to the idea of the existence of missing important explanatory variables. McCullagh and Nelder (1, Section 6.2) describe several scenarios that generalize the Poisson regression formulation and would lead to overdispersion. An important one of these concerns situations in which there is intersubject variability and the Poisson mean for an individual is itself a realization of a random variable. This can result in a more highly variable or heavier-tailed distribution for the response variable, such as the negative binomial (25).

As described by McCullagh and Nelder (1, Section 6.2), when the precise mechanism that produces the overdispersion is unknown, the following algorithm should be employed: parameter estimates are obtained by maximizing the Poisson likelihood, with the estimated variance matrix adjusted by a factor $\hat{\phi}$, defined as the ratio of the deviance to its associated degrees of freedom. Under the more generalized formulations that lead to overdispersion, this method is shown to produce appropriate parameter estimates and inferences, and it has become the conventional approach in Poisson regression analysis. Thus, the presence of a moderate amount of overdispersion, in the absence of evidence of other concerns mentioned previously in this section (e.g., numerous outliers, systematic residual patterns), does not preclude the application of Poisson regression methodology.

4. Sample Size

Signorini (26) is the only paper to date that discusses approximate sample size calculations for Poisson regression, closely paralleling similar calculations for logistic regression. An example from this paper is provided here,

which also serves to illustrate an application in randomized controlled clinical trials for the treatment of epilepsy.

A common response (efficacy) variable in these trials is the number of epileptic seizures, say, per month, assumed to follow a Poisson distribution. Equal allocation of patients to each of the two treatment groups, a follow-up treatment period of 1 month for all patients ($t_i = 1$), a significance level of 0.05, and a power of 0.99 is assumed in the following. Given a seizure rate of 1 per month for the placebo group ($\lambda_{\text{Placebo}} = 1$) and ignoring the explanatory variables, the required total sample size to detect a 50% and 30% reduction in seizure rate for patients on active treatment (corresponding to $\lambda_{\text{Active}} = 0.50$ and $\lambda_{\text{Active}} = 0.70$, respectively) will be 223 and 752 patients, respectively. The total sample size decreases proportionally with the mean follow-up treatment period.

Explanatory variables, such as age, age at onset of epilepsy, baseline number of seizures, type of seizure, and so forth, can be accounted for by assuming for them a distribution that is a member of the multivariate exponential family, with known parameters. Derivations for the special case of normally distributed covariates are illustrated by Signorini (26); the sample size computed without adjusting for covariates is multiplied by the antilog of a quadratic form involving the mean vector and correlation matrix of the covariates and the vector of coefficients from the model that correspond to those covariates. Inclusion of explanatory variables in this manner will reduce the required sample size. In practice, however, dependable estimates of the quantities necessary for this computation may not be available.

If overdispersion is expected, the required total sample size should be increased by a factor ϕ (see Section 3), which must be estimated prior to the calculations. Overdispersion is likely to be present in this application, as noted by Balish *et al.* (27). In addition, as in all clinical trials, the total sample size should be adjusted upward to account for the expected proportion of patients who withdraw from the trial or are "lost to follow-up"; this proportion is generally nonnegligible among patients with epilepsy.

5. Example

In this section, an example is presented demonstrating the use of Poisson regression in the analysis of data from a clinical trial and illustrating many of the issues discussed in previous sections. The data set was presented and described in Byar (28); variables given there that are relevant to the analysis presented below appear in Appendix 1. An analysis of this data set was given by Byar *et al.* (29).

This data set is from a randomized clinical trial conducted by the Veterans Administration Co-operative Urological Research Group. Patients entered the trial with superficial bladder tumors, which were removed. Random assignment was made to three treatments: placebo, pyridoxine pills, or instillation of a chemotherapeutic agent, thiotepa. At subsequent monthly follow-up visits, any tumors noticed were removed and the treatment was continued. The goal of the analysis was to determine the effect of treatment on the frequency of tumor recurrence, defined as the presence of at least one tumor at a follow-up visit. For each of 118 patients (116 with follow-up data), variables on the data set analyzed were: treatment, number of follow-up visits, number of recurrences, number of tumors present at the time of randomization, and diameter (in centimeters) of the largest of these.

The analyses presented in Byar *et al.* (29) used a variety of statistical methods, including survival analysis for the time to first recurrence, F-tests for comparisons of treatment group averages of patient recurrence rates, and χ^2 tests for proportion differences. None of these analyses is fully satisfactory in the sense of employing proper distributional assumptions and having the ability to incorporate all available information (concerning both the magnitude of the response and differences in baseline characteristics). It is reasonable to hypothesize that the number of visits with tumor recurrence might be well approximated by a Poisson distribution. Below we present a Poisson regression analysis of this data set. The analysis was performed using PROC GENMOD in SAS. Details on the use of the procedure are fully explained in the referenced technical report from SAS Institute, Inc. (4), and the specific code that was used in the analysis is presented in Appendix 2.

Table 1 summarizes several aspects of the tumor recurrence data, including numbers of patients in each treatment group with no tumor recurrences, the mean and median values of the patient recurrence rates (# recurrences/# visits), and the overall treatment group rates (total # of recurrences/total # of visits). The different summary measures reflect different aspects of the data. The median was nonzero only for the placebo group, since for the other two treatments, just over half the patients had no recurrences. The mean was particularly high for the pyridoxine group, pre-

Table 1. Summary Statistics for Bladder Cancer Data

Treatment	# of patients	# with 0 recurrences	Mean	Median	Rate
Placebo	47	18	0.053	0.028	0.057
Pyridoxine	31	16	0.075	0	0.057
Thiotepa	38	20	0.044	0	0.038

dominantly due to two patients, each with high rates based on two recurrences in a small number of visits. The overall recurrence rate was lowest for the thiotepa group.

It can be seen that the baseline data might contain important explanatory information: arbitrarily dividing patients according to whether or not they had more than one tumor at baseline, the overall treatment recurrence rates presented in Table 2 suggest that this characteristic impacts future recurrence.

A Poisson regression model was fit to the data, with the canonical (log) link. The response variable was the number of recurrences; explanatory variables included in the model were the number of baseline tumors and size of the largest baseline tumor, as well as the log of the number of visits [from Section 2, this is model (3); as discussed, this is a slightly more general formulation than model (2), which allows the possibility that mean response is not proportional to time].

Overdispersion from the assumed model was indicated by values of 2.43 and 2.42, respectively, for ratios of deviance and Pearson's χ^2 to the associated degrees of freedom. As a result, standard errors and test statistics were adjusted based on the estimated deviance parameter, as described in Section 3.

The fitted equation giving the predicted value of the log of the number of recurrences for each patient was

$$\ln \hat{\mu} = \text{GROUP} + 0.8137 \times \log(\text{months}) + 0.1879 \times (\#\text{ of baseline tumors}) - 0.0087 \times (\text{baseline tumor size}), \quad (6)$$

where GROUP represents a parameter estimate of the effect of each treatment, and which equals -2.5802 for placebo, -2.5488 for pyridoxine, and -3.0955 for thiotepa. The likelihood ratio test comparing the treatment groups yielded a p-value of 0.125; 1 d.f. contrast p-values for pairwise comparisons between the treatments were: 0.906 for placebo versus pyridoxine, 0.069 for placebo versus thiotepa, and 0.081 for pyridoxine versus thiotepa.

Descriptive estimates of the differences between groups can be obtained using the estimated GROUP parameters from Eq. (6). For example, the recurrence rate for thiotepa relative to placebo is $\exp(-3.0955 + 2.5802) =$

Table 2. Recurrence Rates by Baseline Tumor Status

	Overall recurrence rates	
Treatment	1 baseline tumor	>1 baseline tumor
Placebo	0.041	0.076
Pyridoxine	0.046	0.079
Thiotepa	0.021	0.061

0.597; in other words, it would be expected that a patient treated with thiotepa would experience about 40% fewer recurrences than had that patient been treated with placebo.

Baseline number of tumors was significant as an explanatory variable (p-value = 0.002), while size at baseline was not (p-value = 0.900). The estimated coefficient of the log of the number of visits, 0.814, did not differ significantly from 1, indicating no violation of uniformity of recurrence over time.

Consistent with the observed overdispersion, there were some observations with large associated residuals. In particular, for two patients, the deviance residual (r_{Di}) and Pearson residual (r_{Pi}) both exceeded 3 in magnitude: for patient 67, $r_{D67} = 3.79$ and $r_{P67} = 5.31$; for patient 74, $r_{D74} = 3.09$ and $r_{P74} = 4.06$. Both these patients had a large number of tumor recurrences ($y_{67} = 9$, $y_{74} = 8$) but each had only one tumor at baseline, which resulted in underprediction by the model. The magnitudes of the residuals were felt to be consistent with the types of scenarios of McCullagh and Nelder (1) mentioned in Section 3, in which additional variation is due to intersubject variability, and in which inferences from Poisson regression are valid after adjusting for the overdispersion.

Both deviance residuals and Pearson residuals from the fitted model were plotted against the predicted number of tumor recurrences ($\hat{y}_i$), as well as against each of the explanatory variables (plots are not provided). None of these residual plots revealed a systematic pattern that would cast doubt on the adequacy of the model.

References

1. McCullagh P, Nelder JA: *Generalized Linear Models*, 2nd ed. Chapman and Hall, New York, 1989. (1st ed, 1983.)
2. Dobson AJ: *An Introduction to Generalized Linear Models*. Chapman and Hall, New York, 1990.
3. Hinkley DV, Reid N, Snell EJ, Eds: *Statistical Theory and Modelling*. Chapman and Hall, London, 1991.
4. SAS Institute, Inc.: *SAS Technical Report P-243, SAS/STAT Software: The GENMOD Procedure, Release 6.09*. SAS Institute, Inc., Cary, NC, 1993.
5. Aitkin M, Anderson D, Francis B, Hinde J: *Statistical Modelling in GLIM*. Oxford University Press, Oxford, 1989.
6. Kleinbaum DG, Kupper LL, Muller KE: *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. PWS-Kent, Boston, 1988.
7. Myers RH: *Classical and Modern Regression with Applications*, 2nd ed. PWS-Kent, Boston, 1990.
8. Frome EL, Kutner MH, Beauchamp JJ: Regression analysis of Poisson-distributed data. *J Am Statist Assoc* 68:935–940, 1973.

9. Frome EL: The analysis of rates using Poisson regression models. *Biometrics* 39:665–674, 1983.
10. Frome EL, Checkoway H: Use of Poisson regression models in estimating incidence rates and ratios. *Am J Epidemiol* 121:309–323, 1985.
11. Frome EL, Morris MD: Evaluating goodness of fit of Poisson regression models in cohort studies. *Am Statist* 43:144–147, 1989.
12. Gallagher JC, Riggs BL, Recker RR, Goldgar D: The effect of calcitriol on patients with postmenopausal osteoporosis with special reference to fracture frequency. *Proc Soc Exp Biol Med* 191:287–292, 1989.
13. Parker RA: Analysis of surveillance data with Poisson regression: A case study. *Statist Med* 8:285–294, 1989.
14. Vonesh EF: Modelling peritonitis rates and associated risk factors for individuals on continuous ambulatory peritoneal dialysis. *Statist Med* 9:263–271, 1989.
15. Campbell MJ, Machin D, D'Arcangues C: Coping with extra Poisson variability in the analysis of factors influencing vaginal ring expulsions. *Statist Med* 10:241–254, 1991.
16. Rochon J, Aprile MA, Cardella CJ: Analyzing the number of "rejection episodes" in renal transplant studies. *Transplantation* 53:1236–1242, 1992.
17. Neugebauer R, Paik M, Hauser WA, Nadel E, Leppik I, Susser M: Stressful life events and seizure frequency in patients with epilepsy. *Epilepsia* 35:336–343, 1994.
18. Davison AC, Tsai, CL: Regression model diagnostics. *Int Statist Rev* 60:337–353, 1992.
19. Williams DA: Generalized linear model diagnostics using the deviance and single case deletions. *Appl Statist* 36:181–189, 1987.
20. Pierce DA, Schafer DW: Residuals in generalized linear models. *J Am Statist Assoc* 81:977–986, 1986.
21. Davison AC, Gigli A: Deviance residuals and normal score plots. *Biometrika* 76:211–221, 1989.
22. Wang PC: Adding a variable in generalized linear models. *Technometrics* 27:273–276, 1985.
23. Wang PC: Residual plots for detecting nonlinearity in generalized linear models. *Technometrics* 29:435–438, 1987.
24. Lee AH: Diagnostic displays for assessing leverage and influence in generalized linear models. *Aust J Statist* 29:233–243, 1987.
25. Lawless JF: Negative binomial and mixed Poisson regression. *Can J Statist* 15:209–225, 1987.
26. Signorini DF: Sample size for Poisson regression. *Biometrika* 78:446–450, 1991.
27. Balish M, Albert PS, Theodore WH: Seizure frequency in intractable partial epilepsy: A statistical analysis. *Epilepsia* 32:642–649, 1991.
28. Byar DP: Multiple tumor recurrence data for patients with bladder cancer. In: *Data: A Collection of Problems from Many Fields for the Student and Research Worker* (Andrews DF, Herzberg AM, Eds). Springer-Verlag, New York, 1985, pp. 253–259.
29. Byar DP, Blackard C, the VACURG: Comparisons of placebo, pyridoxine, and topical thiotepa in preventing recurrence of stage I bladder cancer. *Urology* 10:556–561, 1977.

Appendix 1. Bladder Cancer Data Set

Patient number	Months of follow-up	Number of recurrences	Recurrence rate	Tumors at baseline	Size of largest baseline tumor
Treatment = Placebo					
1	0	0		1	1
2	1	0	.000	1	3
3	4	0	.000	2	1
4	7	0	.000	1	1
5	10	0	.000	5	1
6	10	1	.100	4	1
7	14	0	.000	1	1
8	18	0	.000	1	1
9	18	1	.056	1	3
10	18	2	.111	1	1
11	23	0	.000	3	3
12	23	2	.087	1	3
13	23	3	.130	1	1
14	23	3	.130	3	1
15	24	4	.167	2	3
16	25	3	.120	1	1
17	26	0	.000	1	2
18	26	1	.038	8	1
19	26	2	.077	1	4
20	28	1	.036	1	2
21	29	0	.000	1	4
22	29	0	.000	1	2
23	29	0	.000	4	1
24	30	2	.067	1	6
25	30	3	.100	1	5
26	30	5	.167	2	1
27	31	3	.097	1	3
28	32	0	.000	1	2
29	34	0	.000	2	1
30	36	0	.000	2	1
31	36	1	.028	3	1
32	37	0	.000	1	2
33	40	4	.100	4	1
34	40	6	.150	5	1
35	41	0	.000	1	2
36	43	1	.023	1	1
37	43	1	.023	2	6
38	44	3	.068	2	1
39	45	5	.111	1	1
40	48	1	.021	1	1
41	49	0	.000	1	3
42	51	1	.020	3	1
43	53	1	.019	1	7

Appendix 1. Continued

Patient number	Months of follow-up	Number of recurrences	Recurrence rate	Tumors at baseline	Size of largest baseline tumor
44	53	5	.094	3	1
45	59	0	.000	1	1
46	61	9	.148	3	2
47	64	5	.078	1	3
48	64	8	.125	2	3

Patient number	Months of follow-up	Number of recurrences	Recurrence rate	Tumors at baseline	Size of largest baseline tumor
Treatment = Pyridoxine					
49	0	0		8	1
50	2	0	.000	1	2
51	4	2	.500	4	6
52	4	0	.000	1	1
53	5	2	.400	1	1
54	7	0	.000	2	3
55	8	0	.000	1	1
56	8	1	.125	4	3
57	11	1	.091	1	1
58	14	0	.000	1	1
59	26	0	.000	1	2
60	29	0	.000	1	2
61	30	1	.033	8	1
62	32	0	.000	1	3
63	33	0	.000	1	1
64	34	5	.147	1	4
65	37	5	.135	3	7
66	38	0	.000	1	1
67	39	9	.231	1	2
68	40	0	.000	1	1
69	40	0	.000	1	1
70	42	9	.214	3	1
71	45	0	.000	1	1
72	45	1	.022	2	1
73	46	2	.043	1	4
74	46	8	.174	1	1
75	48	1	.021	1	1
76	54	0	.000	2	1
77	54	2	.037	1	1
78	55	8	.145	4	1
79	57	0	.000	1	1
80	60	0	.000	3	8

(*continued*)

Appendix 1. Continued

Patient number	Months of follow-up	Number of recurrences	Recurrence rate	Tumors at baseline	Size of largest baseline tumor
Treatment = Thiotepa					
81	1	0	.000	1	3
82	1	0	.000	1	1
83	5	1	.200	8	1
84	9	0	.000	1	2
85	10	0	.000	1	1
86	13	0	.000	1	1
87	14	1	.071	2	6
88	17	5	.294	5	3
89	18	0	.000	5	1
90	18	1	.056	1	3
91	19	1	.053	5	1
92	21	2	.095	1	1
93	22	0	.000	1	1
94	25	0	.000	1	3
95	25	0	.000	1	5
96	25	0	.000	1	1
97	26	3	.115	1	1
98	27	1	.037	1	1
99	29	1	.034	2	1
100	36	2	.056	8	3
101	38	0	.000	1	1
102	39	4	.103	1	1
103	39	7	.179	6	1
104	40	4	.100	3	1
105	41	0	.000	3	2
106	41	0	.000	1	1
107	43	2	.047	1	1
108	44	0	.000	1	1
109	44	5	.114	6	1
110	45	0	.000	1	2
111	46	1	.022	1	4
112	46	0	.000	1	4
113	49	0	.000	3	3
114	50	0	.000	1	1
115	50	3	.060	4	1
116	54	0	.000	3	4
117	54	1	.019	2	1
118	59	0	.000	1	3

Appendix 2. SAS Code

The following lines of code were used to generate the analyses described in Section 5, using SAS Version 6.08:

Appendix 2. Continued

```
PROC GENMOD;
  CLASS TRT;
MODEL N_RECUR =
     TRT N_BASE SIZEBASE LN_MONTH
          DIST = POISSON
          DSCALE
          TYPE3
          OBSTATS;
CONTRAST 'PLA - PYR' TRT 1 -1 0;
CONTRAST 'PLA - THI' TRT 1 0 -1;
CONTRAST 'PYR - THI' TRT 0 1 -1;
MAKE 'OBSTATS' OUT = OUTDATA;
```

Variable names are as follows:

N_RECUR:	number of tumor recurrences
TRT:	treatment group; 1 = placebo, 2 = pyridoxine, 3 = thiotepa
N_BASE:	number of tumors at baseline
SIZEBASE:	size of the largest tumor at baseline
LN_MONTH:	log of the number of monthly follow-up visits

The link function did not need to be specified, since the default was the canonical link corresponding to the DIST option in the model statement, which in this case was the log link. In order to perform the analysis as in model (2), with events proportional to time, LN_MONTH would be omitted from the MODEL statement, and the option OFFSET = LN_MONTH would be added. The option DSCALE adjusts the estimated variance matrix of parameter estimates by the ratio of the deviance to its associated d.f. The option TYPE3 gives Type III statistics (as in PROC GLM) for likelihood ratio tests of each explanatory variable. The option OBSTATS produces a number of statistics associated with each observation and is output to the data set OUTDATA by the MAKE statement. The statistics produced include residuals; RESDEV and RESCHI are the variable names assigned to the deviance residuals and Pearson residuals, respectively.