Published by iambr.info/AMBR

Open Access Original Work

Estimating the effects of the factors underlying the progression of Familial Adenomatous Polyposis using Longitudinal Com-Poisson Model

Vandna Jowaheer[1], Naushad Ali Mamode Khan[1], Durga Charan Pati[2]

To cite: Jowaheer V, Khan NM, Pati DC. Estimating the effects of the factors underlying the progression of Familial Adenomatous Polyposis using Longitudinal Com-Poisson Model. *Arch Med Biomed Res.* 2014;1(1):16-21.

Publication history
Received: December 7, 2013
Revised: March 19, 2014
Accepted: March 21, 2014

Correspondence to
Vandna Jowaheer;
vandnaj@uom.ac.mu

ABSTRACT

This paper aims at developing a statistical model capable of quantifying the effects of various factors on the progression of Familial Adenomatous Polyposis (FAP), a genetic disorder affecting the colon and rectum in human beings. The progression of FAP in affected individuals is monitored by counting the number of polyps developed over a period of time. These count responses repeatedly observed over time are over-dispersed and highly correlated resulting into a complicated longitudinal count data structure, which render the application of commonly, used Gaussian regression model useless. We designed a statistical model based on Com-Poisson distribution, which can efficiently analyze such a data structure. The estimates of over-dispersion as well as correlation parameters confirm the nature of real data. Analysis of the model indicates that males are 50% more at risk to develop polyps than females. With respect to the type of treatment, the application of vitamin C and E with high fiber treatment is a better remedy followed by vitamin C and E only as compared to placebo. In men as well as in women, initial polyp counts positively affect the polyp counts at a given time.

KEY WORDS: ***Familial adenomatous polyposis; Longitudinal responses; Correlation and overdispersion; Com-Poisson model***

INTRODUCTION

Adenomatous polyps are benign pedunculated outgrowths from epithelium with varying malignant potential. It is an autosomal dominant disorder diagnosed when the patient has more than 100 polyps in large bowel or when a member of familial adenomatous polyposis (FAP) family has any number of colonic adenomas detectable. It can also occur in stomach, duodenum and small intestine. The incidence is between 1 in 8000 -28000 individuals. Its main risk is large bowel cancer. The gene responsible is APC gene on the short arm of chromosome 5[1]. It can occur sporadically by new mutation. In these cases large bowel cancer occurs in young adulthood. Male and females are both affected. FAP can be associated with benign tumours like abdominal wall tumours (desmoid tumour) and bone tumours (osteoma).

Polyps are usually visible by the age of 15 years by sigmoidoscopy. Carcinomatous changes (nearly 100%) occur 10-20 years after onset of polyposis. Hence, if left untreated the person develops cancer by the age of 35-40. A person with familial polyposis has a 50 percent chance of passing the condition down to each child. The symptoms associated with the growth of polyps are gastrointestinal problems such as diarrhea, constipation, abdominal cramps, blood in the stool, or weight loss. Patients may also develop other nonmalignant tumors, bone and dental abnormalities. They may also exhibit a spot on the retina of the eye. The patients are usually given a surgery treatment for FAP. It is important to control the recurrence and spread of the polyps in such patients.

The spread of polyposis can be monitored by counting the number of polyps observed in patients over time. Gender, the initial number of polyps at the time of detection and the type of treatment are some of the important factors determining the progression of familial polyposis. One of the two types of treatments: Vitamin C+E and Vitamin C+E+high fibre is usually administered to the patients in order to control the polyp counts. This results into longitudinal count data where the responses are correlated and highly over-dispersed. Moreover, the correlation structure is unknown since the joint distribution of the polyp counts is unknown. Gaussian regression model cannot take into account such features of the polyp counts data and its application will provide highly inefficient estimates of the factor effects.

Stukel[2] and Crouchley and Davis[3] proposed to use the generalized estimating equations (GEE) and random effects modelling approaches respectively to analyze such type of polyp counts data. The 'working' correlation structure based GEE approach by Liang and Zeger[4] suffers from the drawbacks of using misspecified 'working' correlation matrix as highlighted by Crowder[5] and Sutradhar and Das[6]. Hence, the GEE approach assuming an approximate covariance structure based on an equi-correlation structure model used by Stukel[2] fails to yield efficient estimates of the regression parameters. Also, the random effects models similar to those designed by Thall and Vail[7] and Crouchley and Davis[6] are not suitable as they are only able to model the over-dispersion but are not efficient in modelling the time-lag correlations among the counts repeatedly collected over time as discussed by Jowaheer and Sutradhar[8]. Moreover, the estimation of the regression parameters by evaluating integrated likelihood function is quite complicated and the efficiency of these estimates depend on the assumption of the distribution of the random effects. On the contrary, Jowaheer and Sutradhar[8] proposed a negative binomial longitudinal model which models the over-dispersion and used joint generalized estimating equations based on true autocorrelation structure of the count responses repeatedly collected over time and estimated the true parameters involved in the model. Khan and Jowaheer[9] used this negative binomial longitudinal regression model based on stationary autocorrelation structure to analyse polyps data. In this paper, we propose to use a Com-Poisson longitudinal model[10] and use joint generalized quasi-likelihood (GQL) estimating equations[11] based on true stationary autocorrelation structure of the counts to re-analyze the rectal polyps data from Stukel[2].

MATERIALS AND METHODS

Description of the Polyps data

The original data analyzed by Stukel[2] consists of the rectal polyp counts of 58 patients recorded over nine visits along with the information on gender and two baseline measures of the polyp counts taken before the treatment as well as the type of treatment. However, there are some missing counts in these data. In this application, we exclude the patients with

missing data and consider only 45 subjects for 9 three monthly visits. The means and variances of the responses for the 9 visits are shown in **Table 1**. The average lag-correlations are displayed in **Table 2**.

Table 1: Summary statistics of the polyp counts from visit 1 to visit 9

Visits	Sample Mean	Sample Variance
1	6.5778	46.4313
2	7.0444	70.5556
3	7.5556	78.9798
4	8.2444	102.1434
5	8.6889	141.3101
6	8.0667	158.2455
7	7.9556	197.2253
8	7.4000	78.74555
9	6.9778	104.3404

It is noted from **Table 1** that variances are larger than their corresponding means, thus indicating that data are highly overdispersed. Also, the lag-correlations displayed in **Table 2** decrease gradually as the lags increase showing an auto-regressive pattern underlying the responses repeatedly collected over 9 visits. There are three covariates: types of treatment, gender and the sum of base-line rates (BR). Patients are allocated to one of the three-treatment groups-Placebo, Vitamin C+E (TR 1) and Vitamin C+E+high fibre (TR 2). These three groups are represented by x_1 (TR 1), x_2 (TR 2) and placebo being the reference group. Hence, $x_1 = 0$ and $x_2 = 0$ stands for an individual allocated to placebo; $x_1 = 1$ and $x_2 = 0$ stands for an individual allocated to TR 1; $x_1 = 0$ and $x_2 = 1$ stands for an individual allocated to TR 2. The covariate gender is represented by x_3 which is 0 for male and 1 for female. The sum of baseline rates (BR) is the third covariate represented by x_4.

Table 2: Sample correlation values

Legs	Sample correlations
1	0.7881
2	0.6324
3	0.5771
4	0.4821
5	0.3671
6	0.3001
7	0.2775
8	0.2010

Com-Poisson Regression Model

In order to estimate the effect of covariates on the number of polyps developed over a period of 2 years after the start of the treatment, we propose to use a Com-Poisson regression model based on AR (1) type autocorrelation structure[11]. In this section, we provide the structure of this model[10]. The parameters of this model will be estimated using joint generalized quasi-likelihood (JGQL) estimation approach discussed in the next section.

Let y_{it} be a count response and x_{it} be a p-dimensional vector of covariates for subject $i(i = 1,\ldots,I)$ observed at time $t(t = 1,\ldots,T)$. Let β be the $p \times 1$ vector of regression parameters. For the i th subject, let $y_i = \left(y_{i1},\mathrm{K}\ ,y_{it},\mathrm{K}\ ,y_{iT}\right)^T$ be the $T \times 1$ response vector and $X_i = \left(x_{i1},\mathrm{K}\ ,x_{iT}\right)^T$ be the $T \times p$ matrix of covariates. We assume y_{it} follows Com-Poisson distribution[11] with probability mass function.

$$f(y_{it}) = \frac{\lambda_{it}^{y_{it}}}{(y_{it}!)^{\nu}} \frac{1}{Z(\lambda_{it},\nu)} \quad (1)$$

where

$$Z(\lambda_{it},\nu)=\sum_{j=0}^{\infty}\frac{\lambda_{it}^{j}}{(j!)^{\nu}} \quad \& \quad \lambda_{it}=\exp(x_{it}^{T}\beta) \qquad (2)$$

and the parameter ν is the dispersion index such that $\nu=1,\nu<1 \text{ and } \nu>1$ correspond to equi-, over- and under-dispersion. Since equation (1) doesn't have closed form expression, an asymptotic expression[10] is used. This expression is given by:

$$Z(\lambda_{it},\nu);\ \frac{\exp\left(\nu\,\lambda_{it}^{\frac{1}{\nu}}\right)}{\lambda_{it}^{\frac{\nu-1}{2\nu}}(2\pi)^{\frac{\nu-1}{2}}\sqrt{\nu}} \qquad (3)$$

Hence,

$$E(Y_{it})=\theta_{it}=\lambda_{it}^{1/\nu}-\frac{\nu-1}{2\nu};\ Var(Y_{it})=\frac{\lambda_{it}^{1/\nu}}{\nu} \text{ and}$$

$$Cov(Y_{it},Y_{i,t-k})=\rho^{k}\left(\frac{\theta_{it}}{\nu}+\frac{\nu-1}{2\nu^{2}}\right). \qquad (4)$$

Here, $I=45, T=9 \text{ and } p=5$.

Since all the covariates are time-independent,

$$\theta_{i1}=\theta_{i2}=\mathrm{K}\ =\theta_{it}=\mathrm{K}\ =\theta_{i9} \qquad (5)$$

Estimation of Model Parameters

The parameters of the model considered in equations (1) to (4) are estimated using the consistent and efficient joint generalized estimation approach[11], which is briefly explained in this section. The JGQL estimating equation to estimate the regression and over-dispersion parameters is given by:

$$\sum_{i=1}^{I}D_i^{T}\Sigma_i^{-1}(f_i-\mu_i)=0 \qquad (6)$$

where

$f_i=\left(f_{i1}^{T},\mathrm{K}\ ,f_{it}^{T},\mathrm{K}\ ,f_{i9}^{T}\right),\mu_i=\left(\mu_{i1}^{T},\mathrm{K}\ ,\mu_{i1}^{T}\mathrm{K}\ ,\mu_{i1}^{T}\right)$

are $2T\times 1$ vectors with

$f_i=\left(y_{it},y_{it}^{2}\right),\mu_{it}=\left(\theta_{i1},m_{i1}\right)^{T},\theta_{i1}=E(Y_{it})$ and

$m_{i1}=E(Y_{it}^{2})=\frac{\lambda_{i1}^{\frac{1}{\nu}}}{\nu}+\theta_{i1}^{2}$ where

$\theta_{i1}=\exp(x_{i1}^{T}\beta);\Sigma_i$ is the covariance matrix of the score vector f_i and D_i is the $2T\times(p+1)$ derivative matrix consisting of:

$D_i=\left[\partial\mu_i/\partial\beta^{T},\partial\mu_i/\partial\nu\right]=\left[D_{i1}^{T},\mathrm{K}\ ,D_{it}^{T},\mathrm{K}\ D_{i9}^{T}\right]^{T}$,

with $D_{it}=\begin{pmatrix}\partial\theta_{i1}/\partial\beta^{T} & 0\\ \partial m_{i1}/\partial\beta^{T} & \partial m_{i1}/\partial\nu\end{pmatrix}$.

The mathematical details of the covariance matrix Σ_i are available in Mamode Khan and Jowaheer[11]. Note that:

$$\hat{\rho}_l=\frac{\sum_{i=1}^{I}\sum_{t=1}^{9-l}y_{it}y_{it+l}/(9-l)}{\sum_{i=1}^{I}\sum_{t=1}^{9}y_{it}^{2}/9} \qquad (7)$$

for $(l=|t-w|=1,\mathrm{K}\ ,8)$ where $y_{it}=\frac{y_{it}-\theta_{i1}}{\sqrt{\lambda_{i1}^{\frac{1}{\nu}}/\nu}}$. The iterative solution of JGQL estimating equation (6) is given by:

$$\begin{pmatrix}\hat{\beta}_{r+1}\\ \hat{c}_{r+1}\end{pmatrix}=\begin{pmatrix}\hat{\beta}_{r}\\ \hat{c}_{r}\end{pmatrix}+\left[\sum_{i=1}^{I}D_i^{T}\Sigma_i^{-1}D_i\right]_r^{-1}\left[\sum_{i=1}^{I}D_i^{T}\Sigma_i^{-1}(f_i-\mu_i)\right]_r \qquad (8)$$

where $\hat{\beta}_r$ is the value of $\hat{\beta}$ at the r^{th} iteration. $[.]_r$ is the value of the expression at the r^{th} iteration. The data analysis can be easily performed using open-source software R[12].

RESULTS AND DISCUSSION

The results after fitting the model to the data are presented in the **Table 3.**

The average estimates of the correlation parameters are provided in **Table 4**.

The model fits the data very well. The estimates of the lag-autocorrelation values are large, indicating that the data are highly correlated and decreasing values with increasing time lag justifies AR (1) autocorrelation pattern. The treatment parameters TR 1 and TR 2 are both negative, indicating that both the treatments are capable of reducing the number of cancerous polyps when compared to placebo. However, we may conclude that vitamin C and E with high fibre treatment is more effective in the reduction of polyps as compared to vitamin C and E. The negative sign in the sex parameter makes us deduce that there is lesser number of polyps among the female group. The growth of polyps is lesser by almost 50 percent in females as compared to males. Also, if the baseline rates increase by 1 percent, then the

polyp counts will show an increase of 3 percent. The estimate of is significant justifying that the data are over-dispersed. These findings are in line with the findings made by Khan and Jowaheer[10] after fitting an alternative negative-binomial model to the same data set. It should be remarked that Com-Poisson model is preferred to negative-binomial model due to its flexibility of accommodating different types of dispersion structures.

Table 3: Estimates of regression parameters

Variables	Parameters	Estimates	Standard Errors
INTC	β_0	1.2011	0.3010
TR 1	β_1	-0.1111	0.0199
TR 2	β_2	-0.1231	0.0321
Sex	β_3	-0.5021	0.0724
BR	β_4	0.0325	0.0146
Dispersion	ν	0.7821	0.1991

Table 4: Estimates of correlation parameters

Legs	Sample correlations
1	0.7189
2	0.6301
3	0.5712
4	0.4881
5	0.3501
6	0.2901
7	0.2871
8	0.2003

CONCLUSION

Familial polyps, once arising in a human excretory system, multiply quite fast and lead to cancer. The growth of these polyps can be monitored by counting the number of polyps. It is of interest to understand and estimate the effect of important factors such as the type of treatment, sex as well as the baseline counts on the growth of the polyps over time. The polyps count data, longitudinally collected together with the information on covariates, is generally over-dispersed with gradually decreasing auto-correlation pattern. The analysis of such data is quite challenging and requires the application of a properly designed statistical model. In this paper, we have analysed polyps count data using the longitudinal Com-Poisson regression model based on AR (1) type auto-correlation structure. The estimation of the regression and over-dispersion parameters is done using a joint generalized quasi-likelihood approach. The estimates thus obtained are reliable and consistent with very small standard errors. Based on this study, we may conclude that the application of Vitamin C and E with high fibre treatment is a better remedy followed by Vitamin C and E only as compared to placebo in the reduction of polyps in bodies. Males are 50 percent more at risk of developing polyps than females. Hence, with a familial history of polyposis, the offsprings especially the males should be more at guard and must take recourse to early medical check-ups with respect to the disease.

Author affiliations
[1]University of Mauritius, Mauritius
[2]Department of Surgery, SSR Medical College, Mauritius

REFERENCES

1. Kinzler KW, Nilbert MC. Identification of FAP locus genes from chromosome 5q21. *Science*. 1991;253(5020):661-5.
2. Stukel TA. Comparison of methods for the analysis of longitudinal interval count data. *Stat Med*. 1993;12(14):1339-51.
3. Crouchley R, Davies RB. A comparison of population average and random effect models for the analysis of longitudinal count data with base-line information. *J Royal Statist Soc.* 1999;162(3):331-47.
4. Liang KY, Zeger SI. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.

5. Crowder M. On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*. 1995;82:407-10.
6. Sutradhar BC, Das K. On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*. 1999;86:459-65.
7. Thall PF, Vail SC. Some covariance models for longitudinal count data with over-dispersion. *Biometrika*. 1990;46:657-71.
8. Jowaheer V, Sutradhar BC. Analysing longitudinal count data with over-dispersion. *Biometrika*. 2002;89:389-99.
9. Khan NM, Jowaheer V. Analysing familial polyposis using negative binomial longitudinal regression model. Conference proceedings of International Conference on Medical, Biological and Pharmaceutical Sciences, Thailand. 2011.
10. Shmueli G, Minka T, Borle J, Boatwright P. A useful distribution for fitting discrete data. *J Royal Statist Soc*. 2005; 54: 127-42.
11. Khan NM, Jowaheer V. Comparing joint GQL estimation and GMM adaptive estimation in COM-Poisson longitudinal regression model. *Commun Stat-Simul C*. 2013;42(4):755-70.
12. R- development core team. http://www.r-project.org/.