The Report Committee for Hakan Goren

Certifies that this is the approved version of following report:

A FAMILIAL LONGTITUDINAL COUNT DATA STUDY

APPROVED BY

SUPERVISING COMMITTEE

Supervisor:

_____

Daniel A. Powers

_____

Michael Daniels

A Familial Longitudinal Count Data Study

by
Hakan Goren, B.S. Stat.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May 2013

Dedication

This report is dedicated to my parents, my siblings, my nephews and Peter.

Acknowledgements

I would like to thank Dr. Daniel A. Powers and Dr. Michael Daniels who despite of their extremely busy schedule supervised me to complete this report. I also take this opportunity to thank my professors from the SSC department for helping me through. I also would like to thank Vicki Keller for her remarkable help and care for the department's students.

May 2013

ABSTRACT

A Familial Longitudinal Count Data Study

Hakan Goren, M.S.Stat.

The University of Texas at Austin, 2013

Supervisor: Daniel A. Powers

In this report, I study familial longitudinal count data with a Poisson regression model. The data is collected from individuals who are nested in families. I focus on two main issues to fit a model. The first one is the large number of excess zeros and the second one is multi-level random effects. My approach for solving these problems are to use either Zero Inflated Poisson (ZIP) or Negative Binomial (NB) models to control for the excess zeros which allow for estimation of another parameter for over dispersion while developing the model with individual and familial random effects.

First, I use a Poisson regression model with only main effects. After that, I fit a ZIP model to control for the extra zeros. I provide information about general form of the exponential families and a discussion about the dispersion parameter. I also fit a Negative Binomial model instead of the ZIP model. I also build these models with only individual random effects and with both individual and familial random effects as well. I discuss the generalized estimating equation (GEE) approach to estimate the parameters of a generalized linear model with auto regressive correlation between outcomes.

Table of Contents

INTRODUCTION

A typical longitudinal study is a collection of repeated measurements over time from individuals. Count data from individuals over time is commonly obtained in fields such as sociology, epidemiology, and medicine. It is known that repeated measurements correlation provides reduced errors when it is taken into account. Similarly, if the individuals come from families (clusters) then it is natural to assume that the within-family correlation in addition to the repeated measurements. This will provide reduced residual variability as well. Thus, the explained variation gets larger (Burton, Scurrah, 2005).

A Poisson regression model is often the first approach to analyze count data. However, Poisson models have a strong assumption that the variance and the mean of the population is the same. Often time some, unobserved phenomena are involved in the data collection process and cause extra zeros in the data set. In this case there are some zeros from the Poisson distribution and some extra zeros from where the probability of a count is zero for that particular measurement.

If the observations are positively correlated, which often occurs with longitudinal data, then the variances of the time-independent predictor variables (variables that estimate the group effect (or between-subject effect) such as gender, race, treatment, and so on) are underestimated if the data is analyzed as though the observations are independent. In other words, the Type I error rate (rejecting the null hypothesis when it is true, in other words, a false positive) is inflated for these variables (Dunlop, 1994)

Having a larger variance than mean for a Poisson distribution is called over-dispersion. This can happen by either having extra zeros or having extra variance components in the model or having both like this study.

DATA

Health care utilization data for six years from 1985 to 1990 was collected by the Health Science Center, Memorial University, St. John's, Canada (Sutradha, 2011). There are 180 individuals and 48 families in the data. Each individual is nested in a family. Thirty-six families have four members and twelve families have three members. The dependent variable is the "Number of physician visits from 1985 to 1990." Covariates are gender (0=female, 1=male), age (individual's age in 1985), education level (0=high school or less, 1=college or more), chronic disease status (0=no chronic disease, 1=at least one chronic disease). All covariates are time independent.

Denoting the dependent variable by Y; Y_{ijk} represents the observation from the *j*-th member of the *i*-th family in the *k*-th year, where $i = 1,2,3, \dots, m \quad j = 1,2, \dots, \mathrm{n_i} \quad k = 1,2 \dots, 6$. Total number of individuals is $n = \sum_1^m n_i = 180$ and total number of observations is $N = \sum_{i=1}^{m} \sum_{j=1}^{n_i} n_{ij} = 1080$.

DESCRIPTIVE ANALYSIS

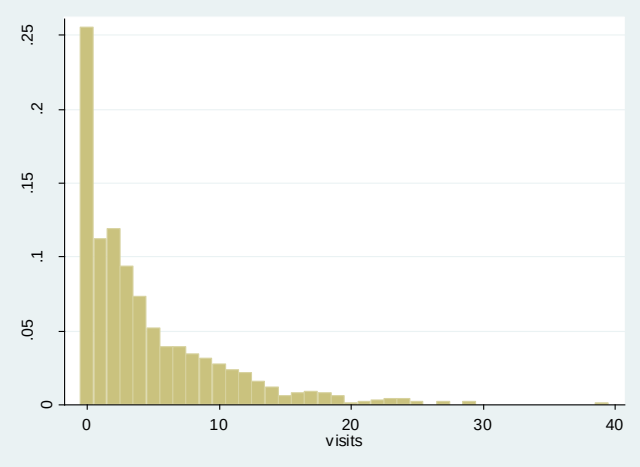
There are 84 female participants, 52 of whom have at least one chronic disease, while 40 out of 96 male participants have at least one chronic disease. The minimum age is 19.9 and maximum age is 85.2 with mean 38.57 years and standard deviation 16.52 years.

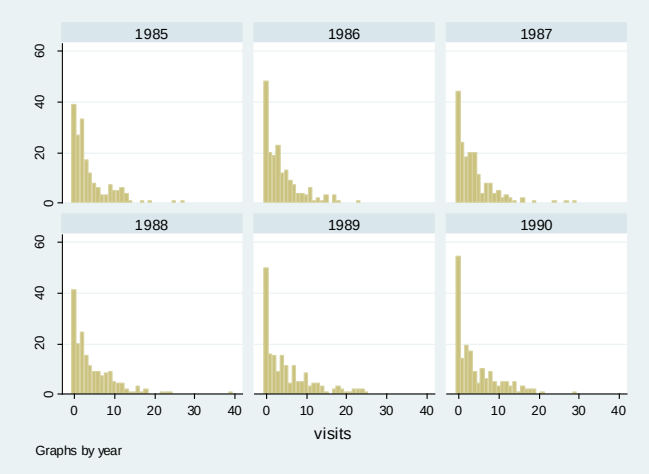
Descriptive analysis shows that the unconditional mean number of visits for all observations is 4.44 while the variance is 27.6. This is evidence for over dispersion. Output and graphs in [Appendix 2] also show that zeros comprise 25% of the all observations. They appear more than expected in the data, with or without conditioning on covariates.

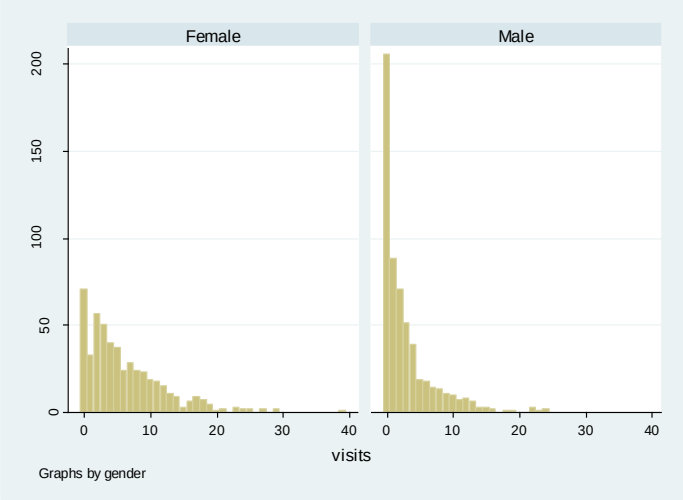
	male		**Female**		
	no disease	**disease**	**no disease**	**disease**	**∑**
HS or less	32	17	19	21	***89***
COL or more	24	23	13	31	***91***
∑	***56***	***40***	***32***	***52***	***180***

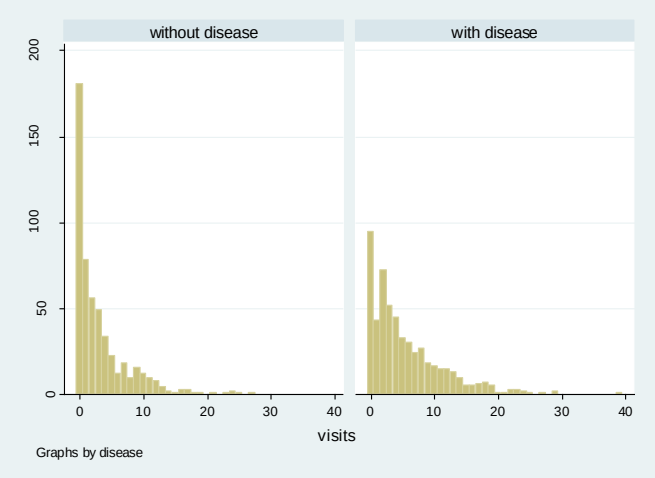
visits	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	276	25.56	276	25.56
1	121	11.2	397	36.76
2	128	11.85	525	48.61
3	101	9.35	626	57.96
4	79	7.31	705	65.28
5	56	5.19	761	70.46
6	42	3.89	803	74.35
7	42	3.89	845	78.24
8	37	3.43	882	81.67
9	34	3.15	916	84.81
10	29	2.69	945	87.5
11	25	2.31	970	89.81
12	23	2.13	993	91.94
13	17	1.57	1010	93.52
14	12	1.11	1022	94.63
>14	58	5.37	1080	100

Table 1 [up]: Number of individuals conditional on the covariates
Table 2 [down]: Frequency table of the number of visits









Graph 1 [up left]: A histogram of all visits
Graph 2 [up right]: A histogram of visits conditional on the year
Graph 3 [down left]: A histogram of visits conditional on gender
Graph 4 [down right]: A histogram of visits conditional on disease

METHODS

The simplest model for count data is a Poisson regression model. Therefore, the first model to fit the data is Poisson regression. For simplicity, let's assume that all the observations are independent and identically distributed Poisson with the parameter λ.

Model 1 - Poisson Regression:

$$f(y_{ijk}|\lambda) = \frac{e^{-\lambda}\,\lambda^{y_{ijk}}}{y_{ijk}!} \qquad y_{ijk} = 0,1,2\ldots;\ \lambda > 0$$

$$\log(\lambda) = \eta = X_{ijk}^{T}\beta \qquad \lambda = e^{\eta} = e^{X_{ijk}^{T}\beta}$$

$$f(y|X,\beta) = \frac{\exp\left(-\exp\left(X_{ijk}^{T}\beta\right)\right)\ .\exp\left(y\left(X_{ijk}^{T}\beta\right)\right)}{y_{ijk}!}$$

Model 2 - Zero Inflated Poisson Regression:

An exponential family is defined with the parameters θ and $\emptyset$ as:

$$f(y) = \exp\left[\frac{y\theta - a(\theta)}{\emptyset} + S(y,\emptyset)\right]$$

Defining $\theta = X^{T}\beta$ $a(\theta) = \exp(\theta)$ $S(y,\emptyset) = -\log(y!)$ $\phi = 1$, we can show that the Poisson distribution is an exponential family. Here $\emptyset$ represents the dispersion parameter and the variance and

the mean of the Poisson distribution is the same when it equals 1. A dispersion parameter greater than 1 indicates over dispersion for the Poisson distribution.

The zero inflated Poisson (ZIP) model is one way to allow for over dispersion caused by extra zeros. This model assumes that the sample is a "mixture" of two sets of individuals: one set whose counts are generated by the standard Poisson regression model, and another set who have zero probability of a count greater than 0. In our study, an empirical way to think about this is that under the assumption $\phi = 1$, which means $E[Y] = Var[Y] = \lambda$, the expected number of zeros can be obtained by substituting $\bar{y}$ as an unbiased estimator of λ in the Poisson probability mass function. Hence, the expected probability of zero is $f(y = 0|\lambda) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-4.44} = 0.0118$. However, more than 25% of all observations are zero in data. Expected number of zeros is much smaller than observed zeros. Therefore, a zip model for this study could be interpreted as a set of individuals who get ill in a given year who have a non-zero probability of seeing the doctor and another set of people who never get ill in that given year and have zero probability of seeing a doctor. Observed values of 0 could come from either group. This suggests building a two-stage model. A logistic model to assign an individual to the set they belong to and a Poisson model for those who belong to the Poisson process. For simplicity, let's assume all observations are independent. Covariate matrix X_{ijk} is assumed to be the same for both states. One can have a different covariate matrix for the zero state if there is a belief about what covariates related to having zero count (Brian H. Neelon , A. James O'Malley and Sharon-Lise T. Normand, 2010).

$$\text{logit}\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \xi_{ijk} = X_{ijk}^T A$$

$$\log(\lambda_{ijk}) = \eta_{ijk} = X_{ijk}^T \beta$$

Let z_{ijk} be an unobserved binary variable indicating if y_{ijk} comes from the latent class zero or non-zero. Decomposition of the complete data log likelihood into two orthogonal components can be obtained by treating the realization of the incidence of extra zeros as a missing latent variable. (Sutradha, 2011)

$$l_C = l_\xi + l_\eta$$

$$l_\xi = \sum_{ijk} \left(z_{ijk} \xi_{ijk} - \log(1 + \exp(\xi_{ijk})) \right)$$

$$l_\eta = \sum_{ijk} \left((1 - z_{ijk}) \left(y_{ijk} \eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!) \right) \right)$$

$$z_{ijk}{}^{(g)} = \begin{cases} \dfrac{1}{1 + \exp\left[-\left(x_{ijk}^T \hat{A}^{(g)} \right) - \exp(x_{ijk}^T \hat{\beta}^{(g)}) \right]} & \text{if } y_{ijk} = 0 \\ 0 & \text{if } y_{ijk} \geq 1 \end{cases}$$

Estimation can be carried out using the EM algorithm. Starting with some initial values for A and β. The EM algorithm proceeds by iteratively replacing z_{ijk} by its conditional expectation $z_{ijk}{}^{(g)}$ where g donates the g-th iteration under the current estimates $\hat{A}^{(g)}, \hat{\beta}^{(g)}$, and solving the likelihood equations of a simpler model Details are given in Appendix 1.

Model – 3 Individual Level Random Effects Longitudinal Data for Poisson Regression

After obtaining Model 1, before we compare it with the ZIP model, let's develop a more complex model. We have data collected from the same individuals over a period of time. It is very likely, and

assumed, that the observations for the same person from one time point to another one will be correlated. Since the time is equally spaced (1 year) for our data and assuming the correlation is stronger when the lag is shorter, an auto-regressive correlation structure is therefore assumed for the variance covariance matrix of v. Let *R* denote the variance covariance matrix of observations with the size of *N*x*N*. *R* is a block diagonal matrix whose blocks are determined by the serial correlation structure of the repeated measurements taken from the same individual. Other elements of the matrix are zero. This applies that individuals are independent from one another.

$$R = \begin{bmatrix} [R_1] & 0 & 0 \\ 0 & [\ddots] & 0 \\ 0 & 0 & [R_n] \end{bmatrix}$$

Therefore, the model can be expressed by,

$$\log(\lambda_{ijk}) = \eta_{ijk} = X_{ijk}^T\beta + v_{ijk} \qquad v \sim N(0, R_v)$$

where R_v is the ***R*** matrix for *v*.

And the log likelihood of the data under this model is

$$l_\eta = \sum_{ijk} \left(y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!) \right) - \frac{1}{2}[N \log(2\pi\sigma_v^2) + v^T R_v^{-1} v]$$

Model – 4 Individual Level Random Effects Longitudinal Data for ZIP Regression

Now let's combine the ZIP model with model 3. So that we will account for excess zeros as well as individual-level random effects.

$$\text{logit}\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \xi_{ijk} = X_{ijk}^T A + s_{ijk}$$

$$\log(\lambda_{ijk}) = \eta_{ijk} = X_{ijk}^T \beta + v_{ijk}$$

$$l_\xi = \sum_{ijk}\left(z_{ijk}\xi_{ijk} - \log(1+\exp(\xi_{ijk}))\right) - \frac{1}{2}[N\log(2\pi\sigma_s^2) + s^T R_s^{-1} s\,]$$

$$l_\eta = \sum_{ijk}\left((1-z_{ijk})\left(y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!)\right)\right) - \frac{1}{2}[N\log(2\pi\sigma_v^2) + v^T R_v^{-1} v]$$

s_{ij} , v_{ij} are random individual effects assumed to be independent and normally distributed with zero mean and variances denoted by R_v and R_s. Selecting this approach leads a Generalized Estimating Equation method. Estimation details are given in the Appendix 1.

$$v \sim N(0, R_v) \qquad s \sim N(0, R_s)$$

where $\boldsymbol{R_v}$, $\boldsymbol{R_s}$ are the ***R*** matrices for *v* and *s* respectively (Andy H. Lee, Kui Wang, 2006).

In statistics, a generalized estimating equation (GEE) can be used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes.

There are several ways to construct the serial correlation from the same individual. A variance component specification, assumes no correlation between repeated measures from the same individual. A compound symmetry specification, assumes non-zero covariance matrix, yet every observation collected from a subject is equally correlated with every other observation from that subject. An auto regressive specification assumes equally spaced time points and stronger correlations for proximate time points than distal time points. If estimation of the regression coefficients is the primary objective of the study, and the number of subjects is much greater than the number of time points, then one should not spend much time choosing a correlation structure. The GEE method for the parameter estimates

was designed to guarantee consistency of the parameter estimates under minimal assumptions about the nature of the time dependence. (Diggle, Liang, and Zeger, 1994)

Model – 4.1: Individual Level Random Effects Longitudinal Data for NB Regression

Negative Binomial (NB) regression is commonly used and recommended when the Poisson data includes too many zeros (Hilbe, 2011). The Poisson distribution has only one parameter for the mean and variance while the NB regression allows an additional parameter for over dispersion. Therefore, we can use the same data to fit a well-known distribution without being worried if the strict assumption of the Poisson regression is met. I will use a Negative Binomial regression with random individual effects instead of a ZIP regression with random individual effects. Therefore, I can still account for extra zeros and random individual effects with a model that is simpler than ZIP regression.

Model – 5: Individual and Familial Level Random Effects Poisson Regression

The following model allows for both individuals and families to have random effects.

$$\log(\lambda_{ijk}) = \eta_{ijk} = X_{ijk}^T\beta + u_i + v_{ijk}$$

$$l_\eta = \sum_{ijk}\Big(\big(1 - z_{ijk}\big)\big(y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!)\big)\Big) - \frac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u^Tu + N\log(2\pi\sigma_v^2) + v^TR_v^{-1}v]$$

$$u \sim N(0, \sigma_u^2) \quad v \sim N(0, R_v)$$

An Extended Model – ZIP Model with Multilevel Random Effects Longitudinal Data:

Unfortunately, this model could not be estimated due to computational difficulties that are discussed in the conclusion section of this paper. However, I will talk about a parameter estimation procedure for this model in Appendix 1. This model is a more general form of other models in this study. So, solutions are similar and simpler with the ease of implementation in SAS.

We have a longitudinal study with individuals nested in families, the model can be written as

$$\text{logit}\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \xi_{ijk} = X_{ijk}^T A + w_i + s_{ijk}$$

$$\log(\lambda_{ijk}) = \eta_{ijk} = X_{ijk}^T \beta + u_i + v_{ijk}$$

$$l_\xi = \sum_{ijk}\left(z_{ijk}\xi_{ijk} - \log\left(1+\exp(\xi_{ijk})\right)\right) - \frac{1}{2}[m\log(2\pi\sigma_w^2) + \sigma_w^{-2}w^T w + N\log(2\pi\sigma_s^2) + s^T R_s^{-1} s\,]$$

$$l_\eta = \sum_{ijk}\left((1-z_{ijk})\left(y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!)\right)\right) - \frac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u^T u + N\log(2\pi\sigma_v^2) + v^T R_v^{-1} v]$$

where w_i and u_i are family random effect, $s_{ij}$and v_{ij} are individual random effects.

$$w \sim N(0,\sigma_w^2) \qquad u \sim N(0,\sigma_u^2) \qquad v \sim N(0,R_v) \qquad s \sim N(0,R_s)$$

where $\boldsymbol{R_v}$, $\boldsymbol{R_s}$ are the ***R*** matrices for *v* and *s* respectively.

ANALYSIS with SAS software and RESULTS

PROC MIXED handles unbalanced data with unequally spaced time points and subjects observed at different time points, uses all the available data in the analysis, directly models the covariance structure, and provides valid standard errors and efficient statistical tests. However, it is not implemented for discrete data. It has been studied extensively. Many options and different structures are available. Random effects and error terms are assumed to be normally distributed with means of 0 and random effects and error terms are independent of each other. The relationship between the response variable and predictor variables is assumed to be linear. Variance-covariance matrices for random effects and error terms exhibit structures available in PROC MIXED.

PROC GENMOD is a procedure in SAS that allows users to run a Poisson regression model for count data. However, it does not allow fitting of a zero inflated model. The model statement supports a choice of an AR (1) error covariance matrix. A random statement is not valid in this context. Repeated option is used instead. This is a powerful tool to conduct Generalized Linear Model regressions as well as the extension to General Estimating Equations where correlated outcome data must be taken into account.

PROC NLMIXED can be viewed as generalizations of the random coefficient models fit by the MIXED procedure. This generalization allows the random coefficients to enter the model nonlinearly, whereas in PROC MIXED they enter linearly. The GLIMMIX procedure also fits mixed models for non-normal data with nonlinearity in the conditional mean function. In contrast to the NLMIXED procedure, PROC GLIMMIX assumes that the model contains a linear predictor that links covariates to the conditional mean of the response. The NLMIXED procedure is designed to handle general conditional

mean functions, whether they contain a linear component or not. As mentioned earlier, the GLIMMIX procedure by default estimates parameters in generalized linear mixed models by pseudo-likelihood techniques, whereas PROC NLMIXED by default performs maximum likelihood estimation by adaptive Gauss-Hermite quadrature. This estimation method is also available with the GLIMMIX procedure (METHOD=QUAD in the PROC GLIMMIX statement).

PROC GLIMMIX fits statistical models to data with correlations or non-constant variability and where the response is not necessarily normally distributed. Conditional on Gaussian random effects, data can have any distribution in the exponential family. It has features like flexible covariance structures for random effects and correlated errors and programmable link and variance functions. *GLIMMIX* uses an iteratively reweighted linear mixed model to estimate a generalized linear mixed model (GLMM). (Wolfinger, R. and O'Connell, M., 1993). This procedure is now fully incorporated into SAS and allows for a number of alternative estimation options.

%GLIMMIX The macro uses iteratively reweighted likelihoods to fit the model. Refer to Wolfinger, R. and O'Connell, M. (1993). By default, %GLIMMIX uses restricted/residual psuedo likelihood (REPL) to find the parameter estimates of the generalized linear mixed model you specify. The macro calls PROC MIXED iteratively until convergence, which is decided using the relative deviation of the variance/covariance parameter estimates. An extra-dispersion scale parameter is estimated by default.

RESULTS Here are the 5 models fit to the data. The first model 'Poisson' is a standard Poisson Regression. The second model 'ZIP' is a zero inflated Poisson Regression. The third model 'Poi nofam' is a Poisson regression model with random individual intercepts. That said, random family effect is ignored and all individuals are treated as independent. The only correlation appears within individual measurements. The fourth model (model 4.1) 'NB nofam' has the same assumptions as the third model except it is a Negative Binomial regression with only individual random effect. This model was preferred to ZIP model since it is easier to code. The main concern is over dispersion, so this regression will include a parameter to relax the assumptions of the Poisson model. The fifth model 'PoiFull' is the full model that is developed in the Method section (model 5). It allows the familial correlation that individuals with in a particular family have in common as well as the longitudinal correlations resulted by the repeated measurements.

To compare the models we need to check fit statistics. Below is a table of model fit statistics that explain how well the data fits the model. This comparison might allow us to find a better model specification.

	Poisson	**ZIP**	**Poi nofam**	**NB nofam**	**PoiFull**
Deviance	5016.6079	6474.57	5016.6079		4234.644
LL	2871.87	3474.424	2871.87		
AIC	7699.6788	6514.57	7699.6788		2941.7
AICC	7699.8846	6515.363	7699.8846		2947.8
BIC	7749.526	6614.264	7749.526		2953.3
QIC			-1089.334	-8231.4533	
QICu			-1109.804	-8248.0933	

Table : Fit statistics obtained from models

LL is the log likelihood fit statistics. It is the likelihood of the data for the given model.

Deviance is a fit statistics obtained by calculating $-2\ (LL_{model} - LL_{full})$. Here full model is a model that there is a parameter for every observation so the data fits perfectly.

AIC is another fit statistics that rewarding the better fit while penalizing for number of parameters (over fitting).

AICC is AIC with correlation for a finite number observations. It has a greater penalty for extra parameters than AIC.

BIC is similar to AIC, penalizes over fitting the model with extra parameters.

To start with, let's check the deviance statistic. The smaller the deviance the better the model. Poisson regression models have smaller deviances than the ZIP model. PoiFull model has the smallest deviance. The concept of the likelihood function does not apply to generalized estimating equations; thus, the usual goodness of fit statistics cannot be computed. Instead, information criteria based on a generalization of the likelihood are computed. The Quasi-likelihood under Independence Model Criterion (QIC) can be used to help choose between two correlation structures, given a set of model terms. The structure that obtains the smaller QIC is "better" according to this criterion. Therefore, there is no deviance statistics calculated for model 'NB nofam'. Checking all other statistics, we can conclude that the PoiFull model performs better than others, as expected.

Parameters	Poisson		ZIP		Poi nofam		NB nofam		PoiFull	
	EST	SE	EST	SE	EST	SE	EST	SE	EST	SE
intercept	1.2583	0.0537	1.5767	0.0554	1.3262	0.2034	1.1696	0.206	1.1886	0.168
y1	-0.1877	0.0509	-0.2807	0.0517	-0.1877	0.0838	-0.1968	0.0913	-0.1877	0.1028
y2	-0.2034	0.0511	-0.2404	0.0518	-0.2034	0.0821	-0.1818	0.0935	-0.2034	0.1012
y3	-0.1779	0.0507	-0.2423	0.0515	-0.1779	0.0882	-0.1183	0.095	-0.1779	0.0965
y4	0.0035	0.0484	-0.0872	0.0489	0.0035	0.0813	0.0514	0.0953	0.003511	0.08446
y5	0.1327	0.0469	0.1138	0.0472	0.1327	0.0674	0.157	0.0715	0.1327	0.06616
age	0.0099	0.001	0.0089	0.001	0.0084	0.004	0.0115	0.0041	0.01167	0.003513
gender	-0.6208	0.0311	-0.3907	0.0315	-0.6872	0.1415	-0.7379	0.1354	-0.6496	0.1026
education	-0.0188	0.0317	0.0337	0.0318	-0.0236	0.1369	-0.036	0.1365	-0.09295	0.1162
disease	0.3059	0.0324	0.1507	0.0325	0.3393	0.127	0.4115	0.1295	0.3414	0.1128
intercept(fam)									0.1151	
AR(1)					0.6034		0.5562		0.5352	
Var									4.2712	

Table: Estimated parameter values from the models

The coefficient for gender (coded 1 for male and 0 for female) shows a negative value for all models. This suggests that females made more visits to the physician compared to males. The positive values of parameter estimates suggest that individuals having at least one chronic disease or belonging to an older age group made more visits to the physicians, as would be expected. The effect on education is not significant in any of the above models.

DISCUSSION and CONCLUSION

Generalized Linear Models provide flexibility to design more and more realistic models to complex data structures. In case of discrete data obtained from a longitudinal study, GEE can be applied to find a solution. The Generalized Estimating Equations procedure extends the generalized linear model to allow for the analysis of repeated measurements or other correlated observations, such as clustered longitudinal data.

In this paper, I tried to develop a model that could better fit data of this nature. To start with, for simplicity, I decided to fit only main effect models. The model, if necessary, can be extended by adding interaction terms. I fit a ZIP model to control for excess zeros to improve the fit of the model. When the software did not allow me to use a more complicated ZIP model, I uses a Negative Binomial regression model to handle over dispersion. Because the data come from a longitudinal study, I built an individual random effects model without the consideration of the familial correlations. I then added the familial random effect. However, due to computational difficulties, I could not estimate the ZIP Model with multi-level random effects. One approach would make it possible for me to solve the extended model by using the *%glimmix* macro with the negative binomial distribution with a log function link. However, the macro works fine only with exponential families. Negative binomial is only an exponential family when the scale parameter *k* is held fixed. Perhaps, defining a gamma function whose parameters are obtained from a negative binomial distribution could lead to an answer. This could be a further study.

Estimates from the models are consistent with expectations. Females, older people, people with at least one chronic disease and people with less education tend to visit a physician more often than their complementary group.

Predicting health care demand can be useful for institutions or governments that provide health care. New methodologies are being developed to understand this phenomenon. The great power of digital computing is the main force behind developing more complex models.

APPENDIX 1

Extended Model Estimation:

$$\text{logit}\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \xi_{ijk} = X_{ijk}^T A + w_i + s_{ijk}$$

$$\log(\lambda_{ijk}) = \eta_{ijk} = X_{ijk}^T \beta + u_i + v_{ijk}$$

A , β correspond to vectors of regression coefficients. w_i and u_i are cluster effects and s_{ijk} , v_{ijk} are individual effects.

$$w = (w_1, w_2, \dots w_m)^T \qquad u = (u_1, u_2, \dots u_m)^T$$

$$s = \left(s_{11}, \dots, s_{1n_1}, s_{21}, \dots s_{2n_2}, \dots, s_{m1}, \dots . s_{mn_m}\right)^T$$

$$v = \left(v_{11}, \dots, v_{1n_1}, v_{21}, \dots v_{2n_2}, \dots, v_{m1}, \dots . v_{mn_m}\right)^T$$

$$\text{logit}\left(\frac{\pi}{1-\pi}\right) = \xi = XA + Q_w w_i + Q_s s_{ijk}$$

$$\log(\lambda) = \eta = X\beta + Q_u u_i + Q_v v_{ijk}$$

A, X, Q_w, Q_s, Q_u, Q_v are design matrices.

$w \sim N(0, \sigma_w^2)$ $\quad u \sim N(0, \sigma_u^2)$ $\quad v \sim N(0, R_v)$ $\quad s \sim N(0, R_s)$

Partition of likelihood: Fixed level and random level

$$l_1 = \sum_{y_{ijk}=0} \log\left(\frac{\exp(\xi_{ijk}) + \exp(-\exp(\eta_{ijk}))}{1 + \exp(\xi_{ijk})}\right)$$

$$+ \sum_{y_{ijk}>0} \left[y_{ijk}\, \eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!) - \log(1 + \exp(\xi_{ijk})\right]$$

$$l_2 = -\frac{1}{2}[m\log(2\pi\sigma_w^2) + \sigma_w^{-2}w^Tw + N\log(2\pi\sigma_s^2) + \mathrm{s}^{\mathrm{T}}\mathrm{R}_s^{-1}s] - \frac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u^Tu + N\log(2\pi\sigma_v^2) + v^TR_v^{-1}v\,]$$

Partition of likelihood: Poisson state and zero state

$$l_\xi = \sum_{ijk}\left(z_{ijk}\xi_{ijk} - \log\left(1+\exp\left(\xi_{ijk}\right)\right)\right) - \frac{1}{2}[m\log(2\pi\sigma_w^2) + \sigma_w^{-2}w^Tw + N\log(2\pi\sigma_s^2) + s^TR_s^{-1}s\,]$$

$$l_\eta = \sum_{ijk}\left(\left(1-z_{ijk}\right)\left(y_{ijk}\eta_{ijk} - \exp\left(\eta_{ijk}\right) - \log\left(y_{ijk}!\right)\right)\right) - \frac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u^Tu + N\log(2\pi\sigma_v^2) + v^TR_v^{-1}v]$$

$$z_{ijk}{}^{(g)} = \begin{cases} \dfrac{1}{1+\exp\left[-\left(x_{ijk}^T\hat{A}^{(g)} - \widehat{w_\iota}^{\,g} - \hat{s}_{ijk}{}^{g}\right) - \exp\left(x_{ijk}^T\hat{\beta}^{(g)}\widehat{u_\iota}^{\,g} - \hat{v}_{ijk}{}^{g}\right)\right]} & \text{if } y_{ijk} = 0 \\ 0 & \text{if } y_{ijk} \geq 1 \end{cases}$$

Derivatives:

$$\frac{dl_\xi}{d\xi} = z - \exp(\xi)/1 + (\exp(\xi)) \qquad \frac{dl_\eta}{d\eta} = (1-z)\left(y - \exp(\eta)\right)$$

$$\frac{d^2l_\xi}{d\xi d\xi^T} = \mathrm{Diag}[-\exp(\xi)/(1+\exp(\xi))^2] \qquad \frac{d^2l_\eta}{d\eta d\eta^T} = \mathrm{Diag}[-(1-z)\exp(\eta)]$$

$$\frac{dl_\eta}{d\beta} = X^T\frac{dl_\eta}{d\eta} \qquad \frac{dl_\eta}{du} = Q_u^T\frac{dl_\eta}{d\eta} - \sigma_u^2u \qquad \frac{dl_\eta}{dv} = Q_v^T\frac{dl_\eta}{d\eta} - \sigma_v^2R_v^{-1}v$$

$$\frac{dl_\xi}{dA} = X^T\frac{dl_\xi}{d\xi} \qquad \frac{dl_\xi}{du} = Q_w^T\frac{dl_\xi}{d\xi} - \sigma_w^2w \qquad \frac{dl_\xi}{ds} = Q_s^T\frac{dl_\xi}{d\xi} - \sigma_s^2R_s^{-1}s$$

$$\Psi_{\alpha ws} = \begin{bmatrix} X^T \\ Q_w^T \\ Q_s^T \end{bmatrix} \left(\frac{d^2 l_\xi}{d\xi d\xi^T} \right) [X \quad Q_w \quad Q_S] + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_w^{-2} I_m & 0 \\ 0 & 0 & R_s^{-1} \end{bmatrix}$$

$$\Psi_{\beta uv} = \begin{bmatrix} X^T \\ Q_u^T \\ Q_v^T \end{bmatrix} \left(\frac{d^2 l_\eta}{d\eta d\eta^T} \right) [X \quad Q_u \quad Q_v] + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_u^{-2} I_m & 0 \\ 0 & 0 & R_v^{-1} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{w} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ w_0 \\ s_0 \end{bmatrix} + \Psi_{\alpha ws}^{-1} \begin{bmatrix} dl_\xi / d\alpha \\ dl_\xi / dw \\ dl_\xi / ds \end{bmatrix} \quad , \quad \begin{bmatrix} \hat{\beta} \\ \hat{u} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ u_0 \\ v_0 \end{bmatrix} + \Psi_{\beta ws}^{-1} \begin{bmatrix} dl_\xi / d\beta \\ dl_\xi / du \\ dl_\xi / dv \end{bmatrix}$$

Hessian Matrix

$$H = \begin{bmatrix} X^T & 0 \\ Q_w^T & 0 \\ Q_s^T & 0 \\ 0 & X^T \\ 0 & Q_u \\ 0 & Q_v \end{bmatrix} + \begin{pmatrix} E\left[-\frac{d^2 l_1}{d\xi d\xi^T} \right] & E\left[-\frac{d^2 l_1}{d\xi d\eta^T} \right] \\ E\left[-\frac{d^2 l_1}{d\eta d\xi^T} \right] & E\left[-\frac{d^2 l_1}{d\eta d\eta^T} \right] \end{pmatrix} + \begin{bmatrix} X & Q_w & Q_s & 0 & 0 & 0 \\ 0 & 0 & 0 & X & Q_u & Q_v \end{bmatrix}$$

Information matrix

$$\Psi_{\alpha ws \beta uv} = H + \begin{bmatrix} 0 & 0 & 0 & & & \\ 0 & \sigma_w^{-2} I_m & 0 & & 0 & \\ 0 & 0 & R_s^{-1} & & & \\ & & & 0 & 0 & 0 \\ & 0 & & 0 & \sigma_u^{-2} I_m & 0 \\ & & & 0 & 0 & R_v^{-1} \end{bmatrix}$$

Information matrix needs to be inverted to get the estimating equations for variances.

Let $V = \left[\Psi_{\alpha ws \beta uv} \right]^{-1}$ $V = [V_{ij}]$ $i = 1,2,3,..,6$ $j = 1,2,3,..6$ block matrices.

$$\sigma_s^2 = \frac{s^T R_s^{-1} s + \text{trace}(R_s^{-1} V_{33})}{N} \qquad \sigma_v^2 = \frac{v^T R_v^{-1} v + \text{trace}(R_v^{-1} V_{66})}{N}$$

$$\sigma_w^2 = \frac{w^T w + \text{trace}(V_{22})}{m} \qquad \sigma_u^2 = \frac{u^T u + \text{trace}(V_{55})}{m}$$

$$SE(\hat{A}) = \sqrt{V_{11}} \qquad SE(\hat{\beta}) = \sqrt{V_{44}}$$

Calculating the auto regressive parameter is possible with a cubic equation (K.K.W. Yau, C.A. McGilchrist, 1998). Let's define three matrices that are symmetric, with *k* (*k*= 6 for this study) rows and columns.

$\boldsymbol{I_i}$ is the identity matrix; $\boldsymbol{J_i}$ has diagonals of one above and below principle diagonal but zero for all other elements; $\boldsymbol{K}_i$ has only two non-zero elements one at each end of the principle diagonal. Recalling variance-covariance structure of individual random effects. If $\boldsymbol{\Psi_i}$ is the block diagonal component of $\boldsymbol{\Psi}$ partitioned conformally to the partition of **R** and then

$$R = \begin{bmatrix} [R_1] & 0 & 0 \\ 0 & [\ddots] & 0 \\ 0 & 0 & [R_n] \end{bmatrix}$$

$\sum_i tr[\boldsymbol{I_i}(\boldsymbol{\Psi_i} + \boldsymbol{R_i})] = L_1 \qquad \sum_i tr[\boldsymbol{J_i}(\boldsymbol{\Psi_i} + \boldsymbol{R_i})] = 2L_2 \qquad \sum_i tr[\boldsymbol{K_i}(\boldsymbol{\Psi_i} + \boldsymbol{R_i})] = L_3$

$$f(\rho) = C_1\rho^3 + C_2\rho^2 + C_3\rho + C_4 = 0$$

$$C_1 = (N-n)(L_1 - L_3) \quad C_2 = (2n-N)L_2 \quad C_3 = NL_3 - (N+n)L_1 \quad C_4 = NL_2$$

With an initial value of the auto-regressive parameter can be estimated by using Newton-Raphson iterative method

$$\hat{\rho} = \rho_0 - [\frac{f(\rho_0)}{f'(\rho_0)}]$$

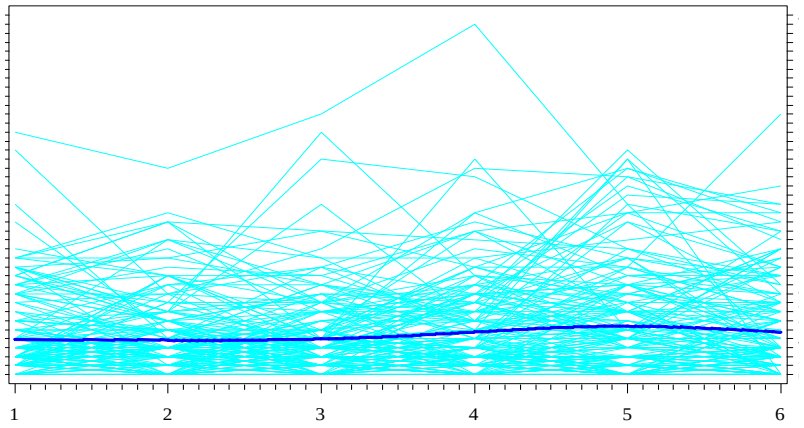
E-M Steps:

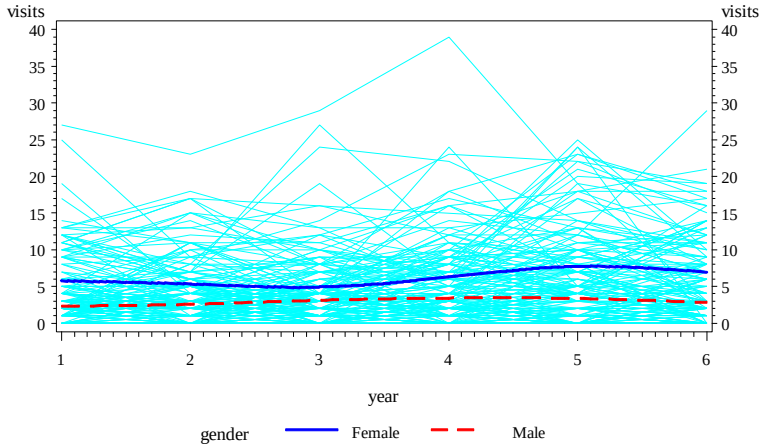
1. Give initial values for $A_0, \beta_0, \ w_0, s_0,$
2. Use fixed variances
3. Calculate z_{ijk}
4. Find ξ and η
5. Calculate the derivatives
6. Obtain new A, β, w, s
7. Go to the third step (do this iteration M times)
8. Get the estimated $\widehat{A}, \ \hat{\beta}, \widehat{w}, \hat{s}$ after M iterations
9. Calculate the first and second derivatives of l_1 with respect to η and ξ
10. Calculate the hessian matrix
11. Calculate the information matrix
12. Get the inverse of the information matrix
13. Get the new values of variance elements
14. Go to the second step and update the variances (do this iteration many times until it converges)

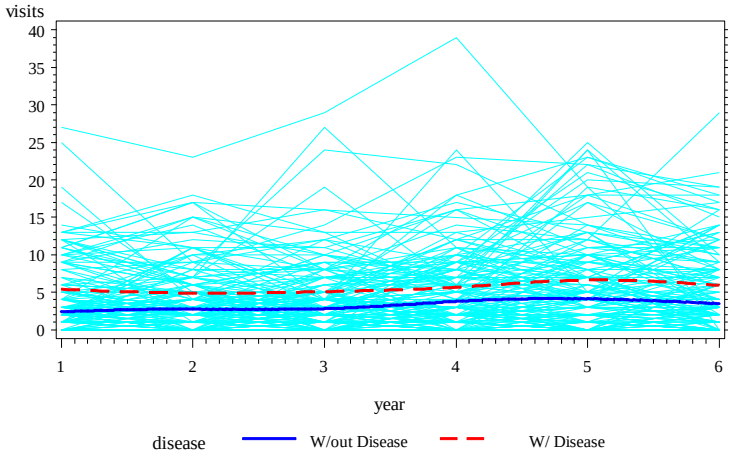
SAS Software CODE FOR MODELS

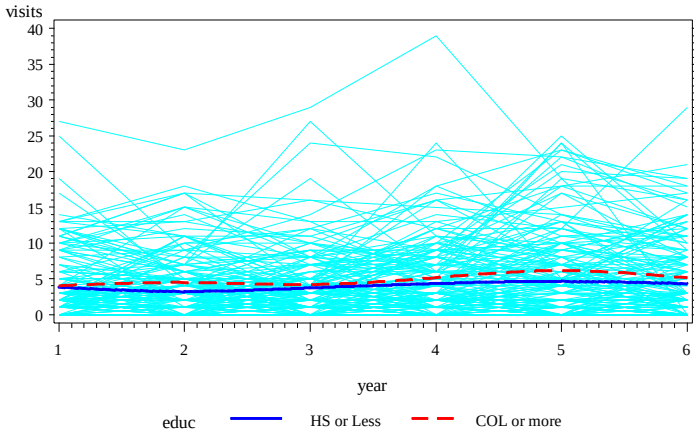
```
/*M1-Basic poisson regression model*/
proc genmod data=hakan.hkn;
class fam id year;
model visits=year age gender disease educ/dist=poisson;
run;

/*M2-a zip regression*/
proc genmod data = hakan.hkn;
class year id fam;
  model visits = year age gender educ disease/ dist=zip ;
  zeromodel year age gender educ disease  /link = logit;
run;

/*M3-individual random effect Poisson Regression*/
proc genmod data=hakan.hkn;
class id year;
model visits=year age gender disease educ/dist=p;
repeated subject=id /corrw covb type=ar(1);
run;

/*M4.1-individual random effect Negative Binomial Regression*/
proc genmod data=hakan.hkn;
class id year;
model visits=year age gender disease educ/dist=nb;
repeated subject=id /corrw covb type=ar(1);
run;


/*M5-multilevel random effect Poisson Regression*/

/*Macro is available online. One has to download and run it before calling
the macro. Since it is pages long I did not put it here. Here is the link for
it:
http://www.stat.ncsu.edu/people/davidian/courses/st762/nlinmix/glmm800.sas */


 %glimmix (data=hakan.hkn,procopt=noclprint,
 stmts=%str(
 class year fam id ;
 model visits= year age gender disease educ;
 random intercept/subject=fam;
 repeated /subject=id(fam)type=ar(1);),
 error=poisson,
 link=log);
run;
```

ividual Profiles with Average Trend Lir









OUTPUT1- Standard Poisson

Model Information	
Data Set	HAKAN.HKN
Distribution	Poisson
Link Function	Log
Dependent Variable	visits

Number of Observations Read	1080
Number of Observations Used	1080

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	1070	5016.6079	4.6884
Scaled Deviance	1070	5016.6079	4.6884
Pearson Chi-Square	1070	5368.2664	5.0171
Scaled Pearson X2	1070	5368.2664	5.0171
Log Likelihood		2871.8700	
Full Log Likelihood		-3839.8394	
AIC (smaller is better)		7699.6788	
AICC (smaller is better)		7699.8846	
BIC (smaller is better)		7749.5260	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.2583	0.0537	1.1531	1.3636	548.94	<.0001
year	1	1	-0.1877	0.0509	-0.2874	-0.0880	13.62	0.0002
year	2	1	-0.2034	0.0511	-0.3035	-0.1033	15.86	<.0001
year	3	1	-0.1779	0.0507	-0.2773	-0.0785	12.30	0.0005
year	4	1	0.0035	0.0484	-0.0913	0.0983	0.01	0.9421
year	5	1	0.1327	0.0469	0.0407	0.2246	8.00	0.0047
year	6	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0099	0.0010	0.0080	0.0118	105.28	<.0001
gender		1	-0.6208	0.0311	-0.6818	-0.5598	397.51	<.0001
disease		1	0.3059	0.0324	0.2423	0.3695	88.90	<.0001
educ		1	-0.0188	0.0317	-0.0810	0.0434	0.35	0.5543
Scale		0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

OUTPUT2- Standard ZIP Regression

Model Information	
Data Set	HAKAN.HKN
Distribution	Zero Inflated Poisson
Link Function	Log
Dependent Variable	visits

Number of Observations Read	1080
Number of Observations Used	1080

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance		6474.5699	
Scaled Deviance		6474.5699	
Pearson Chi-Square	1060	2457.9304	2.3188
Scaled Pearson X2	1060	2457.9304	2.3188
Log Likelihood		3474.4244	
Full Log Likelihood		-3237.2849	
AIC (smaller is better)		6514.5699	
AICC (smaller is better)		6515.3631	
BIC (smaller is better)		6614.2642	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.5767	0.0554	1.4681	1.6853	809.45	<.0001
Year	1	1	-0.2807	0.0517	-0.3820	-0.1794	29.49	<.0001
Year	2	1	-0.2404	0.0518	-0.3420	-0.1388	21.50	<.0001
Year	3	1	-0.2423	0.0515	-0.3432	-0.1414	22.14	<.0001
Year	4	1	-0.0872	0.0489	-0.1831	0.0086	3.18	0.0743
Year	5	1	0.1138	0.0472	0.0213	0.2064	5.81	0.0159
Year	6	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0089	0.0010	0.0069	0.0108	82.02	<.0001
gender		1	-0.3907	0.0315	-0.4525	-0.3289	153.51	<.0001
educ		1	0.0337	0.0318	-0.0286	0.0961	1.12	0.2893
disease		1	0.1507	0.0325	0.0870	0.2144	21.49	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-1.0390	0.2796	-1.5870	-0.4910	13.81	0.0002
year	1	1	-0.5348	0.2641	-1.0524	-0.0172	4.10	0.0429
year	2	1	-0.2135	0.2519	-0.7072	0.2803	0.72	0.3969
year	3	1	-0.3463	0.2562	-0.8484	0.1558	1.83	0.1764
year	4	1	-0.4253	0.2572	-0.9294	0.0787	2.74	0.0981
year	5	1	-0.1104	0.2466	-0.5937	0.3729	0.20	0.6544
year	6	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	-0.0061	0.0053	-0.0166	0.0043	1.31	0.2518
gender		1	1.0417	0.1627	0.7228	1.3606	40.99	<.0001

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
educ		1	0.2358	0.1630	-0.0838	0.5553	2.09	0.1482
disease		1	-0.7215	0.1624	-1.0397	-0.4032	19.74	<.0001

OUTPUT 3 Poisson Individual Random Effect

Model Information	
Data Set	HAKAN.HKN
Distribution	Poisson
Link Function	Log
Dependent Variable	visits

Number of Observations Read	1080
Number of Observations Used	1080

Algorithm converged.

GEE Model Information	
Correlation Structure	AR(1)
Subject Effect	id (180 levels)
Number of Clusters	180
Correlation Matrix Dimension	6
Maximum Cluster Size	6
Minimum Cluster Size	6

Algorithm converged.

Working Correlation Matrix						
	Col1	**Col2**	**Col3**	**Col4**	**Col5**	**Col6**
Row1	1.0000	0.6034	0.3641	0.2197	0.1326	0.0800
Row2	0.6034	1.0000	0.6034	0.3641	0.2197	0.1326
Row3	0.3641	0.6034	1.0000	0.6034	0.3641	0.2197
Row4	0.2197	0.3641	0.6034	1.0000	0.6034	0.3641
Row5	0.1326	0.2197	0.3641	0.6034	1.0000	0.6034
Row6	0.0800	0.1326	0.2197	0.3641	0.6034	1.0000

GEE Fit Criteria	
QIC	-1089.3344
QICu	-1109.8042

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept		1.3262	0.2034	0.9275	1.7249	6.52	<.0001
year	1	-0.1877	0.0838	-0.3520	-0.0235	-2.24	0.0251
year	2	-0.2034	0.0821	-0.3643	-0.0425	-2.48	0.0132
year	3	-0.1779	0.0882	-0.3508	-0.0050	-2.02	0.0438
year	4	0.0035	0.0813	-0.1558	0.1628	0.04	0.9655
year	5	0.1327	0.0674	0.0006	0.2647	1.97	0.0490
year	6	0.0000	0.0000	0.0000	0.0000	.	.
age		0.0084	0.0040	0.0005	0.0163	2.09	0.0365
gender		-0.6872	0.1415	-0.9647	-0.4098	-4.86	<.0001
disease		0.3393	0.1270	0.0903	0.5882	2.67	0.0076
educ		-0.0236	0.1369	-0.2919	0.2448	-0.17	0.8634

OUTPUT 4- NB Individual Random Effect

Model Information	
Data Set	HAKAN.HKN
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	visits

Number of Observations Read	1080
Number of Observations Used	1080

Algorithm converged.

GEE Model Information	
Correlation Structure	AR(1)
Subject Effect	id (180 levels)
Number of Clusters	180
Correlation Matrix Dimension	6
Maximum Cluster Size	6
Minimum Cluster Size	6

Algorithm converged.

Working Correlation Matrix						
	Col1	**Col2**	**Col3**	**Col4**	**Col5**	**Col6**
Row1	1.0000	0.5562	0.3093	0.1720	0.0957	0.0532
Row2	0.5562	1.0000	0.5562	0.3093	0.1720	0.0957
Row3	0.3093	0.5562	1.0000	0.5562	0.3093	0.1720
Row4	0.1720	0.3093	0.5562	1.0000	0.5562	0.3093
Row5	0.0957	0.1720	0.3093	0.5562	1.0000	0.5562
Row6	0.0532	0.0957	0.1720	0.3093	0.5562	1.0000

GEE Fit Criteria	
QIC	-8231.4533
QICu	-8248.0933

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept		1.1696	0.2060	0.7659	1.5734	5.68	<.0001
year	1	-0.1968	0.0913	-0.3757	-0.0180	-2.16	0.0310
year	2	-0.1818	0.0935	-0.3651	0.0015	-1.94	0.0520
year	3	-0.1183	0.0950	-0.3045	0.0678	-1.25	0.2128
year	4	0.0514	0.0953	-0.1354	0.2383	0.54	0.5895
year	5	0.1570	0.0715	0.0169	0.2971	2.20	0.0281
year	6	0.0000	0.0000	0.0000	0.0000	.	.
age		0.0115	0.0041	0.0035	0.0195	2.83	0.0047
gender		-0.7379	0.1354	-1.0032	-0.4725	-5.45	<.0001
disease		0.4115	0.1295	0.1576	0.6654	3.18	0.0015
educ		-0.0360	0.1365	-0.3036	0.2315	-0.26	0.7919

Model Information	
Data Set	WORK._DS
Dependent Variable	_z
Weight Variable	_w
Covariance Structures	Variance Components, Autoregressive
Subject Effects	fam, id(fam)
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Dimensions	
Covariance Parameters	3
Columns in X	11
Columns in Z Per Subject	1
Subjects	48
Max Obs Per Subject	24

Number of Observations	
Number of Observations Read	1080
Number of Observations Used	1080
Number of Observations Not Used	0

Parameter Search					
CovP1	**CovP2**	**CovP3**	**Variance**	**Res Log Like**	**-2 Res Log Like**
0.1151	0.5352	4.2712	4.2712	-1470.8659	2941.7317

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
1	1	2941.73173533	0.00000000

Convergence criteria met.

Covariance Parameter Estimates		
Cov Parm	**Subject**	**Estimate**
Intercept	fam	0.1151
AR(1)	id(fam)	0.5352
Residual		4.2712

Fit Statistics	
-2 Res Log Likelihood	2941.7
AIC (smaller is better)	2947.7
AICC (smaller is better)	2947.8
BIC (smaller is better)	2953.3

PARMS Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	0.00	1.0000

Solution for Fixed Effects						
Effect	**year**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept		1.1886	0.1680	47	7.07	<.0001
year	1	-0.1877	0.1028	1023	-1.83	0.0681
year	2	-0.2034	0.1012	1023	-2.01	0.0446
year	3	-0.1779	0.09650	1023	-1.84	0.0656
year	4	0.003511	0.08446	1023	0.04	0.9669
year	5	0.1327	0.06616	1023	2.01	0.0452
year	6	0	.	.	.	.
age		0.01167	0.003513	1023	3.32	0.0009
gender		-0.6496	0.1026	1023	-6.33	<.0001
disease		0.3414	0.1128	1023	3.03	0.0025
educ		-0.09295	0.1162	1023	-0.80	0.4239

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
year	5	1023	3.49	0.0039
age	1	1023	11.03	0.0009
gender	1	1023	40.05	<.0001
disease	1	1023	9.16	0.0025
educ	1	1023	0.64	0.4239

Description	Value
Deviance	4234.6441
Scaled Deviance	991.4324
Pearson Chi-Square	4348.8109
Scaled Pearson Chi-Square	1018.1616
Extra-Dispersion Scale	4.2712

References

Andy H. Lee, Kui Wang. (2006). Multilevel Zero Inflated Poisson Regression Modelling of Correlated Count Data with Excess Zeros. *Statistical methods in Medical Researches*.

Bernard Kolman, David Hill. (2000). *Elementary Lineer Algebra.* Prentice Hall.

Brian H. Neelon , A. James O'Malley and Sharon-Lise T. Normand. (2010). *A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use.* Sage.

Burton, Scurrah. (2005). Covariance components models for longitudinal family data.

Casella G, Berger R. (2002). *Statistical Inference.* Pacific Grove: Duxbury.

Diggle, Liang, and Zeger. (1994). *Analysis of Longitudinal Data.* New York: Oxford University Press.

Dunlop. (1994). Regression for Longitudinal Data: A Bridge from Least Squares Regression.

E. Barros, J. Achar, J. Mazucheli. (2010). Longitudinal Poisson Modelling : an application for CD4 counting in HIV-infected patiens. *Journal of Applied Statistics*.

Frees, E. W. (2004). *Longitudinal and Panel Data.* Cambridge: Cambridge University Press.

Hardin, James, Hilbe, Joseph. (2003). *Generalized Estimating Equations.* London: Chapman and Hall.

Hasan, M. T. (n.d.). Longitudinal Models for Non-Stationary Exponential Data.

Hilbe, J. M. (2011). *Negative Binomial Regression.* Newyork: Cambridge.

K.K.W. Yau, C.A. McGilchrist. (1998). ML and Reml Estimation in Survival Analysis With Time Dependent Correlated Frailty. *Statistics in Medicine*.

Kung-Yee Liang, Scott Zeger. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 13-22.

Kutner M, Neter J, Li W. (2005). *Applied Linear Statistical Models.* New York: McGraw-Hill/Irwin.

Molenberghs G, Verbeke Geert. (2005). *Models for DiscreteLongitudinal Data.* New York: Springer.

Sutradha, B. (2011). *Dynamic Mixed Models for Familial Longitudinal Data.* New York: Springer.