Estimating social network size using overdispersed Poisson regression model with covariates

BY LI SONG, CHUN YIP YAU
Department of Statistics, Columbia University,
New York City, NY10025, U.S.A.

SUMMARY

The study of networks, sets of objects connected by relationship, is an important area in sociology. It helps to understand the causes and consequences of the structure or relationships in large groups of people. Recently, Zheng et al (2006) used a multilevel overdispersed Poisson regression model to estimate the variation of propensity to form ties to people in some groups (e.g. males in prison, the homeless). This paper extends this model by incorporating covariates to explain the number of acquaintances. This gives more precise estimates of the variation of propensity. Some key words: Overdispersion; Poisson Regression model; Social networks; Social structure; Survey.

1. INTRODUCTION

From the conception of Blau (1974), social structure, or the structure of social networks, is the difference in affiliation patterns from what would be observed if people formed friendships entirely randomly. Studies of social networks have been found to have important implications for many aspects, e.g. the social mobility (Lin 1999), getting a job (Granovetter 1995) and the spread of infectious disease (Morris and Kretzschmar 1995).

One type of survey measuring social network structure asks the questions of the format “How many people do you know named Nicole?”, “How many people do you know incarcerated in prison?”. The responses show a wide range of variation. By examination of the pattern of the variation, we can learn about important characteristic of social network, as well as the processes that create this network.

Recently, Zheng et al (2006) used these “How many X’s do you know?” count data to analyze the social structure of the acquaintanceship network in the United States. More specifically, they fit a multilevel Poisson regression with variance components corresponding to survey respondents and subpopulations and an overdispersion factor that varies by group. Using multilevel model, the individual and subpopulation effects on propensity to form ties can be analyzed separately. Moreover, the variation of the propensity among groups can be estimated by the overdispersion parameter in the model. In this paper, we extend the model by incorporating covariates (e.g. age, sex)
to explain more of the variation of the responses. Therefore, a more precise estimates of the variation of the propensity (overdispersion) can be obtained. In section 2, the data set and the model proposed by Zheng et al (2006) are reviewed. Then our extension of the model is discussed in section 3. Results of our model are included in section 4. Finally, discussion and conclusion are in section 5.

2. Data and the Overpersion Poisson Regression model

The data is from a survey mentioned in McCarty et al (2001). The data are responses from 1370 adults in the United States to a series of questions of the form “How many people do you know in group X?”. 32 such kind of questions were asked for each individual. In addition, background demographic information, including sex, age, income and marital status, was also collected. The responses are truncated at 30.

Zheng et al (2006) proposed a Multilevel Poisson Regression model to reveal information about social structure in the acquaintanceship network. The model is described as

\[ y_{ik} \sim \text{Poisson}(e^{\alpha_i + \beta_k + \tilde{\gamma}_{ik}}) \quad i = 1, 2, \ldots, 1370, \quad k = 1, 2, \ldots, 32 \]  

(2.1)

where \( y_{ik} \) is the number of persons in group \( k \) known by person \( i \), \( \alpha_i \) and \( \beta_k \) are the main effects of individual \( i \) and subpopulation \( k \) respectively, \( \tilde{\gamma}_{ik} \) is the interaction effects induced by individual \( i \) and subpopulation \( k \). Here \( e^{\alpha_i} \) can be interpreted as the expected number of persons known by respondent \( i \), \( e^{\beta_k} \) is the proportion of subgroup \( k \) in the social network, and \( e^{\tilde{\gamma}_{ik}} \) can be interpreted as the relative propensity of individual \( i \) to know a person in group \( k \). For each subpopulation \( k \), we let the multiplicative factor \( e^{\tilde{\gamma}_{ik}} \) follow a gamma distribution with a value of 1 for the mean and a value of \( 1/(\omega_k - 1) \) for the shape parameter. The mean of the gamma distribution is fixed to be 1 to avoid redundance with a location shift in \( \beta_k \). The \( \omega_k \)'s are called the overdispersion parameters because of the following reason. When \( \omega_k = 1 \), the shape parameter in the gamma distribution is infinity, which implies that the relatively propensity \( e^{\tilde{\gamma}_{ik}} \) is a constant 1. On the other hand, the larger the value of \( \omega_k \), the larger the value is the variance of the gamma distribution, meaning that the relatively propensity varies a lot.

This distribution is convenient because the \( \tilde{\gamma} \)'s can be integrated out of (1) to yield the model

\[ y_{ik} \sim \text{Negative - Binomial} (\text{mean} = e^{\alpha_i + \beta_k}, \text{overdispersion} = \omega_k), \]  

(2.2)

with probability density function
\[ f_{y_{ik}}(y|\alpha_i, \beta_k, \omega_k) = \left( y + \xi_{ik} - 1 \right) \left( 1 - \frac{\xi_{ik}}{\omega_k} \right) ^{y} \]

where \( \xi_{ik} = e^{\alpha_i + \beta_k}/(\omega_k - 1) \). This model is estimated with a hierarchical (multilevel) model and Bayesian inference (see, e.g. Gelman et al. 2003). The procedure is simpler than that of our extended model which will be discussed in details in the next section so is not repeated here.

3. Extended Overpersion Poisson Regression model with covariates

3.1. The model

The Overdispersed Poisson Regression proposed by Zheng et al. (2006) provided a basic framework to analyze the “How many X do you know?” type count data. However, the propensity for knowing someone in a certain group is not completely random, but is well dependent on the age, sex and other characteristics of an individual. For example, sales tend to meet more people than housewife, adult have larger social network than children. Therefore, including covariates in the model can help to explain the individual propensity and variation in propensity of forming social network. By virtue of the commonly used generalized linear models (see, e.g. McCullagh and Nelder 1989), it is natural to extend model (2.2) to

\[ y_{ik} \sim \text{Negative - Binomial(} \text{mean} = e^{\alpha_i + \beta_k + \eta_i + \gamma_{ik}}, \text{overdispersion} = \omega_k) \],

(3.1)

where

\[
\begin{align*}
\eta_i & = C_{\alpha} + X_{1i}^{(0)} \phi_1 + X_{2i}^{(0)} \phi_2 + \ldots + X_{mi}^{(0)} \phi_m \\
\gamma_{ik} & = C_{\psi_k} + X_{1i}^{(k)} \psi_{1k} + X_{2i}^{(k)} \psi_{2k} + \ldots + X_{ni}^{(k)} \psi_{nk},
\end{align*}
\]

where \( i = 1, 2, \ldots, 1370, k = 1, 2, \ldots, 32, y_{ik}, \alpha_i, \beta_j, \omega_{ik} \) are the same as in model (2.1), \( X_{ba}^{(c)} \) represents individual \( a \)'s \( b \) – \( th \) covariate among those covariates explaining the effect of group \( c \) (with \( c = 0 \) stands for main effect), \( \phi \)'s and \( \psi \)'s are the corresponding coefficients, \( C_{\alpha} \) and \( C_{\psi_k} \)'s are normalization constants that are to be described in the next subsection. Note that we allow different covariates for each group in the model.

3.2. Identifiability issues

The given model has nonidentifiability. Firstly, any constant \( C \) can be added to all of the \( \alpha_i \)'s and subtracted from all of the \( \beta_j \)'s and the likelihood will remain unchanged because it depends on these parameters only through sums of the form
\[ C = C_1 + \frac{1}{2} C_2, \]

where \( C_1 = \log(\sum_{k \in G_1} e^{\beta_k/P_{G_1}}) \) adjusts for the rare girls’ names and \( C_2 = \log(\sum_{k \in B_2} e^{\beta_k/P_{B_2}}) - \log(\sum_{k \in G_2} e^{\beta_k/P_{G_2}}) \) represents the difference between boys’ and girls’ names. Then \( C \) is added to all of \( \alpha_i \)'s and subtracted from all of the \( \beta_k \)'s. We refer the details and the definition of \( P_{B_1}, P_{G_1} \) and \( P_{G_2} \) to Zheng et al (2006).

Moreover, by the same reason of setting the mean of \( \tilde{\gamma}_{ik} \) equals 1, we choose \( C_\alpha \) and \( C_\psi_k \) such that

\[
E\left( e^{C_\alpha + X_{i1}^{(0)} \phi_1 + X_{2i}^{(0)} \phi_2 + \ldots + X_{mi}^{(0)} \phi_m} \right) = 1, \quad \text{and} \quad E\left( e^{C_\psi_k + X_{i1}^{(k)} \psi_{1k} + X_{2i}^{(k)} \psi_{2k} + \ldots + X_{nki}^{(k)} \psi_{nkk}} \right) = 1, \quad k = 1, 2, \ldots, 32,
\]

to avoid redundancy with a location shift in \( \beta_k \). In practice, we approximate the expectations by the sample means. That is, setting

\[
C_\alpha = - \log \left( \frac{1}{N} \sum_{i=1}^{N} e^{X_{i1}^{(0)} \phi_1 + X_{2i}^{(0)} \phi_2 + \ldots + X_{mi}^{(0)} \phi_m} \right), \quad \text{and} \quad C_\psi_k = - \log \left( \frac{1}{N} \sum_{i=1}^{N} e^{X_{i1}^{(k)} \psi_{1k} + X_{2i}^{(k)} \psi_{2k} + \ldots + X_{nki}^{(k)} \psi_{nkk}} \right), \quad k = 1, 2, \ldots, 32.
\]

\( N \) is the number of respondents in the survey. Apart from that, there will be identifiability problem if the a covariate appears in all the equations involving \( \eta_i \) and \( \gamma_{ik} \). Specifically, if a covariate \( y \) appears in all model equations, then \( \phi \) and \( \psi_k \) varies in the likelihood of each \( y_{ik} \) only in the form \( \phi y_i + \psi_k y_i \). The likelihood is unchanged if a constant \( C \) is added to \( \phi \) and subtracted from all the \( \psi \)'s. If this is the case, we set \( \phi = 0 \) to avoid the unidentifiability issue.

### 3.3. Fitting Algorithm

The parameters \( \alpha_i \)'s, \( \beta_k \)'s and \( \omega_k \)'s, \( \phi_i \)'s and \( \psi_{ik} \)'s are estimated with a hierarchical (multilevel model) and Bayesian inference (see, e.g. Gelman et al. 2003). The respondent parameters \( \alpha_i \)'s are assumed to follow a normal distribution unknown mean \( \mu_\alpha \) and standard deviation \( \sigma_\alpha \). Similarly, the group effect \( \beta_k \) follows a normal distribution \( N(\mu_\beta, \sigma_\beta^2) \). For simplicity, we assign independent uniform \((0,1)\) prior distributions to the overdispersion parameters on the inverse scale--prior \( 1/\omega_k \propto 1 \). The overdispersion parameter \( \omega_k \) are constrained to the range \((1, \infty)\), and so it is convenient
to put a model on the inverses $1/\omega_k$, which fall in $(0,1)$. For the regression coefficients, we assume $p(\psi_{lk}) \propto 1$ and $\phi_l \sim N(\mu_\phi, \sigma^2_\phi)$. The hyperparameters $\mu_\alpha, \mu_\beta, \mu_\phi, \sigma_\alpha, \sigma_\beta, \sigma_\phi$ are estimated from the data with noninformative uniform prior distribution. The joint posterior density can be written as

$$p(\alpha, \beta, \omega, \phi, \psi, \mu_\alpha, \mu_\beta, \mu_\phi, \sigma_\alpha, \sigma_\beta, \sigma_\phi | y) \propto \prod_{i=1}^n \prod_{k=1}^K \left( y_{ik} + \xi_{ik} - 1 \right) \left( \frac{1}{\omega_k} \right)^{\xi_{ik}} \left( \frac{\omega_k - 1}{\omega_k} \right)^{y_{ik}} \times \prod_{i=1}^n N(\alpha_i | \mu_\alpha, \sigma^2_\alpha) \prod_{k=1}^K N(\beta_i | \mu_\beta, \sigma^2_\beta) \prod_{l=1}^L N(\phi_l | \mu_\phi, \sigma^2_\phi).$$

The posterior simulations for the model are obtained by a Gibbs-Metropolis algorithm, iterating the following steps:

1. For each $i$, update $\alpha_i$ using Metropolis with jumping distribution $\alpha_i^* \sim N(\alpha_i^{(t-1)}, \text{jump}_{\alpha_i}^2)$.
2. For each $k$, update $\beta_k$ using Metropolis with jumping distribution $\beta_k^* \sim N(\beta_k^{(t-1)}, \text{jump}_{\beta_k}^2)$.
3. For each $j$, update $\phi_j$ using Metropolis with jumping distribution $\phi_j^* \sim N(\phi_j^{(t-1)}, \text{jump}_{\phi_j}^2)$.
4. For each $l, k$, update $\psi_{lk}$ using Metropolis with jumping distribution $\psi_{lk}^* \sim N(\psi_{lk}^{(t-1)}, \text{jump}_{\psi_{lk}}^2)$.
5. Use Gibbs to update $\mu_\alpha, \sigma^2_\alpha, \mu_\beta, \sigma^2_\beta, \mu_\phi, \sigma^2_\phi$ respectively.
6. For each $k$, update $\omega_k$ using Metropolis with jumping distribution $\omega_k^* \sim N(\omega_k^{(t-1)}, \text{jump}_{\omega_k}^2)$.
7. Renormalizing at the end of each iteration.

### 4. Results

#### 4.1. Models Fitted

To select suitable covariates to be included in the model, i.e., to choose how many $\phi$’s and $\psi$’s to be included, we fitted and compared five models:

1. No covariate. (Null model)
2. All covariates included. (Full model)
3. Use covariates chosen by analysing the residuals of null model. (Pre-selected model)
4. With insignificant coefficients removed after fitting the Full model. (Post-selected from full)

5. With insignificant coefficients removed after fitting the pre-selected model. (Post-selected from pre-selected)

In the pre-selection process, we define $\tilde{y}_i = \sum_{k=1}^{K} y_{ik}$ and define the residuals of null model to be $r_{ik} = \sqrt{y_{ik}} - \sqrt{e^{\alpha_i + \beta_k}}$. Then we choose $\phi$’s by regressing $\tilde{y}_i$’s against all the covariates. Stepwise AIC procedure is used to choose the important covariates to enter the $\phi$ part of the model. In choosing $\psi$’s, for each $k$, we regress $r_{ik}$’s against all the covariates. Then stepwise AIC procedure to choose the important covariates to enter the $\psi$ part of the model.

In Bayesian framework, AIC and BIC are not usually used for model selection because there is a problem in determining the number of parameters due to the use of hyperparameter. Instead, DIC (Gelman, 2006) is suggested as an alternative for AIC and BIC. After fitting the model, we compare the model using DIC criteria, which is defined as

$$DIC = \text{mean}(-2 \log(\text{likelihood})) + 2\text{var}\left(\frac{\text{deviance}}{2}\right).$$

The values and the orders of the DIC are listed as follows after relocating such that the smallest DIC value is 0. Similar to AIC and BIC, the smaller the DIC value, the better the model.

<table>
<thead>
<tr>
<th>Models</th>
<th>Relative DIC</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model</td>
<td>958.67</td>
<td>4</td>
</tr>
<tr>
<td>All covariates (full) model</td>
<td>2072.8</td>
<td>5</td>
</tr>
<tr>
<td>Pre-selected model</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Post-selected model from full model</td>
<td>716.05</td>
<td>3</td>
</tr>
<tr>
<td>Post-selected model from pre-selected model</td>
<td>422.46</td>
<td>2</td>
</tr>
</tbody>
</table>

From the table, we observed the following facts. First, the null model has a relatively high DIC, indicating that the covariates indeed helps to improve the fitting. Second, the Pre-selected model gives the lowest DIC. This implies that the regression of residuals against covariates give satisfactory results. Finally, post-selected models do not give a better fit. One reason is that too many coefficients are removed in the post-selection process. A stepwise procedure may be a better alternative. However, this require much more computational efforts. Therefore we choose the pre-selected model as our final model.

4.2. Coefficients Estimates
As incorporating covariates in the model to explain the overdispersion is our main concern, we only report the coefficient and overdispersion parameter estimates $\phi$, $psi$ and $\omega$. The values of individual and group effects $\alpha$ and $\beta$ will not be reported. Instead, they are compared with the null model in section 4.4.

The regression coefficients $\phi$’s are

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Edu</td>
<td></td>
</tr>
<tr>
<td>Neighbors</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td></td>
</tr>
<tr>
<td>Org</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td></td>
</tr>
<tr>
<td>Democ</td>
<td></td>
</tr>
<tr>
<td>Indep</td>
<td></td>
</tr>
<tr>
<td>Low.inc</td>
<td></td>
</tr>
</tbody>
</table>
The regression coefficients $\psi$'s are as follows. Each graph represents the coefficient $\psi$ of a covariate for each of the 32 groups. A single dot located exactly at point 0 means that the covariate is not included in the model equation of that group.
4.3. **Overdispersion Estimates**

The estimated overdispersion are shown in the following graph. To look at the
overdispersion features more clearly, the group are plotted separately according to low overdispersion (less than 2, left) and high overdispersion (greater than 2, right).

Similar to the result of Zheng et al. (2006), the groups of common used names such as "Michael" and "Christina" have low overdispersion. On the other hand, the group "Jaycee" and "homeless" show high value of overdispersion.

4.4. **Comparision to the null model**

The main goal of this paper is to make use of covariates to improve the fitting of the overdispersed Poisson regression model. Improvement will be indicated by a reduction of overdispersion compare to the null model. Moreover, the effect of individuals $\alpha_i$ and groups $\beta_k$ of the two models should be comparable. To check the agreements of $\alpha_i$’s and $\beta_k$’s, we have the following scatter plot.
The points are scattered around a 45 degrees straight line implies that $\alpha_i$’s and $\beta_k$’s of the two models are comparable. Note that $\alpha_i$’s in the null model tend to be greater than that of the pre-selected model as some of the effects are explained by the $\phi$’s.

The same scatter plot is made to compare the overdispersion parameter $\omega_k$. 

All the points are little under the 45 degrees line meaning that the overdispersion roughly agrees with the null model while having a little reduction. This confirms that the covariates really have explanatory power.

5. Conclusion

This paper successfully extended the overdispersed Poisson regression model proposed by Zheng et al. (2006) by including covariates in the analysis. While the model with covariates gives similar individual and group effects as the null model, overdispersion is reduced. It helps the understanding of the structure of social network by discovering important factors that affect the propensity of knowing people in certain groups.

Acknowledgement

This work was partially supported by the Student Stipends, Columbia University in the City of New York, U.S.A. We would like to thank Prof. Tian Zheng, Miss. Yuejing Ding and Mr. Tyler McCorMick for insightful comments, which substantially improve the paper.

References


