Multilevel modelling of tabular counts via Deconstructed Maximum Likelihood

Jarod Lee*
University of Technology Sydney, Sydney, Australia Yan.Lee@uts.edu.au

James Brown
University of Technology Sydney, Sydney, Australia James.Brown@uts.edu.au

Louise Ryan
University of Technology Sydney, Sydney, Australia Louise.M.Ryan@uts.edu.au

Abstract

A multilevel Poisson regression model is proposed for modelling spatially clustered tabular count data. Counts in tabular data are usually perturbed to reduce disclosure risk, and this can have great impact on small cell counts. We propose a Poisson mixed model with multiplicative random effect that is Gamma distributed. Compared to the existing model using log-Normal random effect, our model utilizes summary statistics of individual with events, which has benefits in terms of privacy protection, data quality improvement and small area analysis, since perturbation of event counts is no longer required. The parameters in our model are estimated via Deconstructed Maximum Likelihood, a new framework where dataset of fundamental different nature, namely individual and area level variables in our context, are used for inference without having to combine them into a single analysis file. The modularity approach is representative of the structure of multilevel data and is scalable as it avoids the repetition of area level variables over individuals living in the common geographical region. Also, our model has closed form likelihood, rendering inference exact without the use of numerical integration. When the variance of the log-Normal random effect is small, the two models produce similar estimates. We apply our model to the Australian unemployment data and show the potential benefits of incorporating it into data extraction tools.

Keywords: Perturbation; Sufficient statistics; Privacy; Modularity.

1. Introduction

Count data is prevalent in wide ranging applications, from unemployment rates in economics, disease mapping in public health to poverty and crime rate in social science. However, unit record data are usually not available for dissemination due to data cust responsibility to protect the privacy of data provider, legally and ethically. This is especially true in health and social sciences where data contain sensitive personal information. Even when personal identifiers such as name, address, date of birth and Medicare number have been removed, linking certain data fields with external datasets can often reveal the identity of individuals with low uncertainty, especially when the data contain time of events and sensitive information such as income and health status. Data breaching, whether intentional or unintentional can result in legal and financial penalty, as well as loss of trust. As such, although it is sometimes possible to release unit record data, the process is elaborate, lengthy and comes with a set of restrictions such as in-house analysis.

That being said, data custodians are usually happy to release cross-classified data in tabular form, where counts are subjected to perturbation (random adjustment) to minimize disclosure risk. Perturbation results in large relative percentage error for small cell counts, thus no reliance should be placed on these cells. It also limits the geographical level of analysis and the number of cross-classifications. As we move to a lower geographical region or as the number of cross-classification increases, the counts become smaller and perturbation has a greater impact on the quality of the data. There is 2 Statistical Disclosure Control (Hundepool et al., 2012) that deal with the balancing of providing users with data of reasonable quality and the need to protect the confidentiality of respondents.

In this paper, we show that perturbation of event counts can be avoided with the clever use of summary statistics. This is useful as perturbation affects event counts more heavily compared to population counts, especially for rare events. Moreover, information regarding individuals and the area where they live can be extracted without t 2 J
S R .J S R 4 2
individual and area level variables come from different sources. DML is not a single statistical method, but rather a conceptual framework for combining datasets of different nature. Here, we consider the Poisson-Gamma model for characterizing geographical variations of events, in which we will describe in Section 2.

2. The model

The mean outcome of individual i in area j, λ_{ij} depends on individual level variables $\boldsymbol{x}_{ij}$ and an area specific random effect b_j via $\lambda_{ij} = b_j \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the regression coefficient vector. The random effect b_j captures the deviation of area specific rates from the overall mean, and is modelled as coming from a Gamma distribution with mean μ_j that depends on area level variables $\boldsymbol{u}_j$ via $\mu_j = \exp(\boldsymbol{u}_j^T \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is the regression coefficient vector. Given the random effect b_j, the outcome y_{ij} is Poisson distributed with mean λ_{ij}, i.e. $y_{ij} \mid b_j \sim \text{Pois}\left(\lambda_{ij}\right)$. The model simultaneously account for the homogeneity of individuals living within the same geographical region and the heterogeneity due to varying individual characteristics.

The log-likelihood of the proposed model can be derived as:

$$\ell(\boldsymbol{\alpha}, \gamma, \kappa; \boldsymbol{y}) \propto \sum_j \left(\sum_{i: y_{ij}=1} \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} \right)$$
$$+ \sum_j \left\{ -\frac{\mu_j}{\kappa} \log(\kappa) - \log\Gamma\left(\frac{\mu_j}{\kappa}\right) + \log\Gamma\left(\omega_j + \frac{\mu_j}{\kappa}\right) - \left(\omega_j + \frac{\mu_j}{\kappa}\right) \log\left[\sum_{i=1}^{n_j} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} + \frac{1}{\kappa}\right] \right\}$$

Individuals with event contribute to the log-likelihood only via the first term, aside from ω_j, number of individuals with event in area j. Using matrix and vector notations, the first term can be written as $\boldsymbol{1}^T \boldsymbol{X}_{event} \boldsymbol{\alpha}$, where $\boldsymbol{X}_{event}$ is the design matrix corresponding to individuals with event. This observation is important since the full dataset containing individuals with event is no longer required for statistical analysis. Instead, the model only requires some summary statistics of individuals with event.

Aside from ω_j, the second terms involve two other quantities from data: $\sum_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})$ which involves a summation over the population at risk in area j; and $\mu_j = \exp(\boldsymbol{u}_j^T \boldsymbol{\gamma})$ that is defined only at the area level. These information can be acquired independently from the datasets containing the population at risk and area level variables respectively. Thus, our model has the capability to analyze data of different nature directly without the need to combine them into a single file.

Figure 1 shows the information flow of the iterative scheme for obtaining maximum likelihood estimates via the Newton Raphson algorithm. For the event dataset, summary statistics are fed into the algorithm once. For both datasets on population at risk and group characteristics, information are fed into the algorithm using initial values for the parameters. The algorithm then version of the parameters, in which the information are fed into the algorithm again using the new parameters. This process is repeated until the algorithm converges.

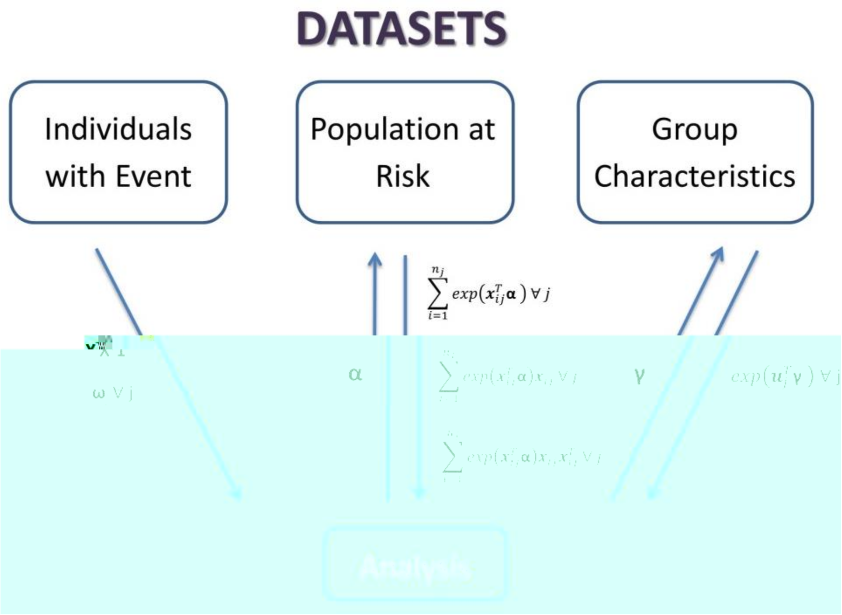

DATASETS
Individuals with Event
Population at Risk
Group Characteristics
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha}) \forall j$
ω ∀ j
α
$\sum_{i=1}^{n_j} exp(x_{ij}^T\alpha)x_{ij} \forall j$
$\sum_{i=1}^{n_j} exp(x_{ij}^T\alpha)x_{ij}x_{ij}^T \forall j$
γ
$exp(u_j^T\gamma) \forall j$
Analysis

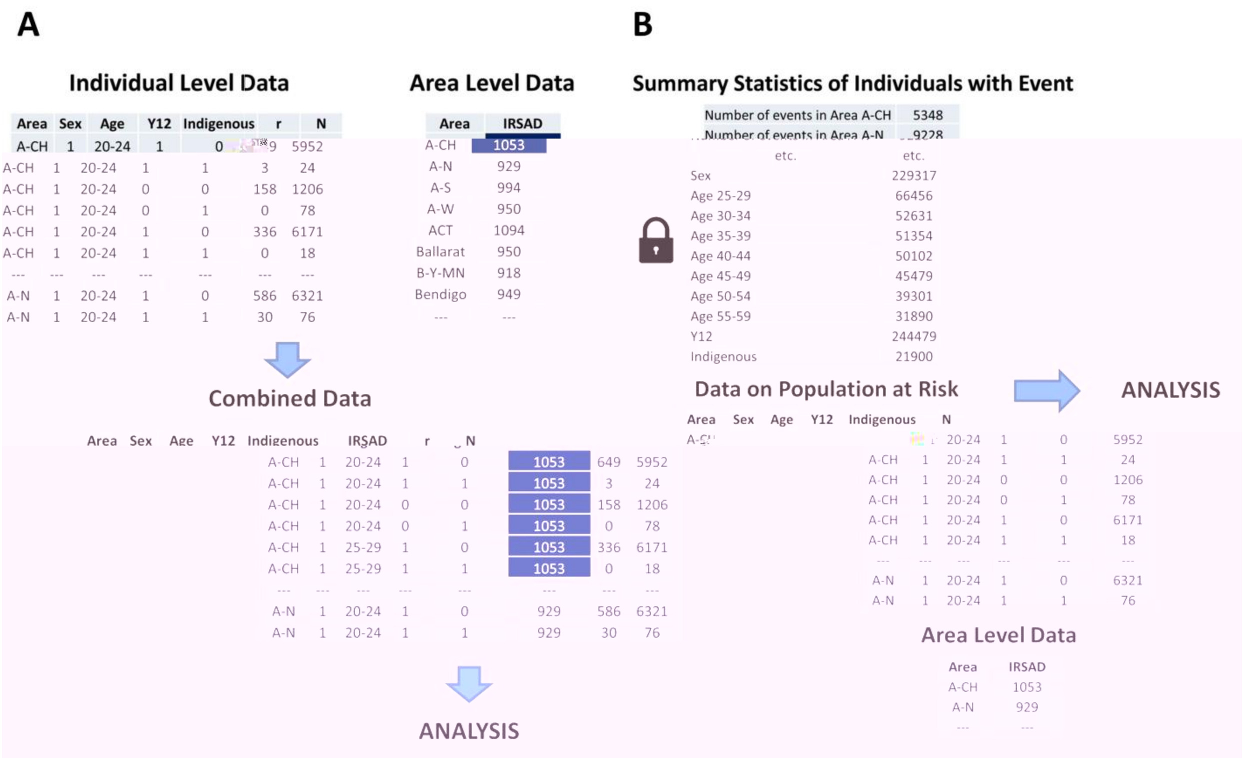
Figure 1. Information flow for the Poisson-Gamma model fitted using the Newton-Raphson algorithm.

Existing model for fitting multilevel count data imposes the log-Normal distribution with mean 0 and variance σ^2 on b_j. This results in intractable likelihood that needs to be solved using approximation (e.g. Laplace approximation) or numerical integration (e.g. Gauss-Hermite quadrature), and the need to combine the individual and group level variables into a single file prior to analysis. In contrast, the integral in the likelihood of our model can be solved exactly. Our model also allows direct input of individual and group level variables, avoiding the repetition of group level variables over individuals living in the same geographical region. For existing model, the marginal expectation is $E(y) = \exp(X\beta + U\gamma + \frac{\sigma^2}{2})$; whereas for our model, $E(y) = \exp(X\beta + U\gamma)$. When σ^2 is small, $\exp(\frac{\sigma^2}{2}) \approx 1$ and thus we should expect the estimates produced by the two models to be similar. Ideally a good model should include as many informative group level variables as possible so that the unexplained variation σ^2 is small.

3. The Australian Unemployment Data

Our research question is to investigate whether area with higher socio-economic advantage is more resilient to Global Financial Crisis (GFC) (2007-8) in terms of unemployment. Socio-economic advantage is represented by the area level variable Index of Relative Socio-economic Advantage and Disadvantage (IRSAD), which is publicly available on the Australian Bureau of Statistics (ABS) website. IRSAD is an index that that ranks areas in Australia according to relative socio-economic advantage and disadvantage. High values of IRSAD indicate high social advantage, and vice-versa. As we want to investigate the effect of socio-economic advantage pre-GFC on the unemployment rates post-GFC, we use the values from 2006 Census, the latest available data before GFC.

Dataset containing individual level variables sex, age, Year 12 completion and indigenous is obtained from the TableBuilder under the 2011 Census. TableBuilder is an online data extraction tool provided by ABS, whereby users can build cross-classification of census variables for geographical areas as defined in the Australian Statistical Geography Standard (ASGS) (ABS, 2012), except Mesh Blocks. We choose Statistical Area Level 4 (SA4) as it is originally designed for the output of labor force survey. As with other tabular data, the counts are perturbed. Also, geographic classification of Australia has changed from Australian Standard Geographical Classification (ASGC) (ABS, 2011) in 2006 to ASGS in 2011. The area level data in 2006 are classified according to Collection District (CD), in which we have to average across multiple CD to match the geographical classification in 2011 (SA4), according to the matching file provided by ABS.



Individual Level Data

Area	Sex	Age	Y12	Indigenous	r	N
A-CH	1	20-24	1	0	[illegible]	5952
A-CH	1	20-24	1	1	3	24
A-CH	1	20-24	0	0	158	1206
A-CH	1	20-24	0	1	0	78
A-CH	1	20-24	1	0	336	6171
A-CH	1	20-24	1	1	0	18
---	---	---	---	---	---	---
A-N	1	20-24	1	0	586	6321
A-N	1	20-24	1	1	30	76

Area Level Data

Area	IRSAD
A-CH	1053
A-N	929
A-S	994
A-W	950
ACT	1094
Ballarat	950
B-Y-MN	918
Bendigo	949
---	---

Combined Data

Area	Sex	Age	Y12	Indigenous	IRSAD	r	N
A-CH	1	20-24	1	0	1053	649	5952
A-CH	1	20-24	1	1	1053	3	24
A-CH	1	20-24	0	0	1053	158	1206
A-CH	1	20-24	0	1	1053	0	78
A-CH	1	25-29	1	0	1053	336	6171
A-CH	1	25-29	1	1	1053	0	18
---	---	---	---	---	---	---	---
A-N	1	20-24	1	0	929	586	6321
A-N	1	20-24	1	1	929	30	76

Summary Statistics of Individuals with Event

Number of events in Area A-CH	5348
Number of events in Area A-N	9228
etc.	etc.
Sex	229317
Age 25-29	66456
Age 30-34	52631
Age 35-39	51354
Age 40-44	50102
Age 45-49	45479
Age 50-54	39301
Age 55-59	31890
Y12	244479
Indigenous	21900

Data on Population at Risk

Area	Sex	Age	Y12	Indigenous	N
A-CH	[illegible]	20-24	1	0	5952
A-CH	1	20-24	1	1	24
A-CH	1	20-24	0	0	1206
A-CH	1	20-24	0	1	78
A-CH	1	20-24	1	0	6171
A-CH	1	20-24	1	1	18
---	---	---	---	---	---
A-N	1	20-24	1	0	6321
A-N	1	20-24	1	1	76

Area Level Data

Area	IRSAD
A-CH	1053
A-N	929
---	---

Figure 2. Data requirement of (A) existing software packages for fitting multilevel models (B) DML.

The unemployment data represents the typical structure of multilevel data, where individual and area level data come from disparate sources. Dataset containing individual level variables are typically available via census, survey or centrally organized disease registries such as Australian Association of Cancer Registries and hospital records; whereas dataset on area level variables are often available via a different database from survey or census. Existing software packages for fitting multilevel models require these datasets to be combined into a single file before performing analysis. This results in the unnecessary repetition of area level variables as indicated by the shaded region in Figure 2A. There could potentially be many area level variables, and existing methods are inefficient. Using the DML framework, these datasets are analyzed directly (Figure 2B). In addition, only summary statistics of individuals with event are needed, which has benefit in terms of privacy protection, data quality improvement and small area analysis.

4. Discussion and Conclusion

The unemployment data are fitted using the proposed model, and we compare the results with the existing Poisson model fitted on the combined data using the *lme4* (Douglas Bates et al., 2014) package within the statistical software R (R Core Team, 2014). The results are summarized in **Table 1**. The estimates and standard errors produced by the Poisson model and DML are very similar, which is expected as σ^2 is small. Both models also agree in terms of the statistical significance of variables.

	Poisson		DML	
Parameter	Est	SE	Est	SE
α_o (Intercept)	-2.06*	0.028	-2.03*	0.027
α_1 (Female)		*Baseline*		
α_1 (Male)	-0.06*	0.003	-0.06*	0.003
α_2 (a20to24)		*Baseline*		
α_2 (a25to29)	-0.51*	0.005	-0.51*	0.005
α_3 (a30to34)	-0.70*	0.005	-0.70*	0.005
α_4 (a35to39)	-0.80*	0.005	-0.80*	0.005
α_5 (a40to44)	-0.92*	0.006	-0.92*	0.006
α_6 (a45to49)	-1.03*	0.006	-1.03*	0.006
α_7 (a50to54)	-1.11*	0.006	-1.11*	0.006
α_8 (a55to59)	-1.10*	0.007	-1.10*	0.007
α_9 (Not completed Year 12)		*Baseline*		
α_9 (Completed Year 12)	-0.46*	0.003	-0.46*	0.003
α_{10} (Not Indigenous)		*Baseline*		
α_{10} (Indigenous)	[illegible]	[illegible]	[illegible]	[illegible]
γ (RSAD)	-0.04	0.027	-0.04	0.027
σ^2	0.07	N/A	N/A	N/A
κ	N/A	N/A	0.06*	0.010

*Significant at $p < 0.001$.

Table 1. Estimates and standard error of the unemployment data, fitted using the existing Poisson model and DML. Est and SE corresponds to the estimates and standard errors respectively.

The unemployment rate for male is 5.82% lower than female (z = -10.00, P < 0.001). Also, an older person has a lower unemployment rate compare to a younger person, with the exception of age category 55-59. For example, the unemployment rate of a person in the age category 25-29 is 39.95% lower compare to a person in the baseline age category 20-24 (z = -102.00, P < 0.001). This is sensible as a person gain more qualifications and experiences as they grow older. However, the unemployment rate of a person in the age category of 55-59 is slightly higher compared to a person in the age category of significant portion of Australian will suffer from unemployment due to insufficient expertise, especially when the world is transiting rapidly from labor-intensive economy to knowledge economy. As expected, the unemployment rate of indigenous Australian is 115.98% higher than non-indigenous Australian (z = 110.00, P < 0.001).

Areas with high level of socio-economic advantage in 2006 have lower unemployment rate in 2011, i.e. more robust to the GFC, albeit the non-significance of the parameter (z = -1.48, P = 0.138). This might be due to the fact that SA4 is the largest sub-State regions among the Main Structure classifications within ASGS, thus the homogeneity of IRSAD values between regions. IRSAD values are obtained by averaging the values of all the sub-regions within the SA4, resulting in the balancing out between low and high values. The use of a lower geographical classification such as SA3 will reduce the homogeneity of IRSAD values, and might result in significant estimate. Ideally, SA3 is preferred over SA4 for a more detail analysis compared to the labor force survey. However, when we move from a higher geographical classification (SA4) to a lower one (SA3), a greater adjustment is needed in order to preserve privacy due to smaller population at risk. Even with SA4, we have approximately 25% of 0 counts and 30% of counts that are lesser than or equal to 5. Perturbation of small cell counts results in large relative percentage error and subsequently affects the quality of data.

Privacy is also a major concern in the data dissemination process. In our model, only the summary statistics of individuals with event is required, as opposed to the whole event data. For the unemployment data, these summary statistics are the number of unemployed in each areas, number of unemployed males, number of unemployed individuals aged 25 to 29, 30 to 34 etc., number of unemployed who completed Year 12 and number of unemployed who are indigenous. Privacy is protected to a certain degree since it is almost impossible to reconstruct information of unemployed individuals based on these summary statistics only. For instance, one cannot tell that person A from area X , who is an aged 35 to 39 indigenous male that completed Year 12, is unemployed by using just the summary statistics. Therefore, data custodians can increase the limit of geographical level in which they can release their data by using our model, allowing detail analysis at a smaller area level. Also, the quality of data can be improved since the summary statistics of event counts can be released without perturbation. Perturbation of population counts is still required, but the effect is relatively small compared to event counts.

Data quality issue has prevented our analysis to be done at the SA3 level. In our analysis, the full event dataset is extracted from TableBuilder before the summary statistics are computed. If our model were to be incorporated into TableBuilder, summary statistics can be extracted directly and thus analysis at SA3 level will be possible.

References

ABS Catalogue No 1216.0.15.002

ABS

Catalogue No 1270.0.55.006

R package version 1.1-7

Statistical disclosure control

R

Foundation for Statistical Computing

The

Guardian