GETTING THE MOST FROM TABULAR DATA

JAROD LEE (UTS)

Prof James Brown (UTS) & Prof Louise Ryan (UTS)

Workshop hosted by UoW – 8th Feb 2017

School of Mathematical and Physical Sciences

OUTLINE

- **REMINDER:** we can do a lot with tabular data.

- Extending to fitting a mixed / random intercept model:
  - Classic approach
  - Alternative approach

- Example with 2011 Australian Census data

FLEXIBLE TABLES

- Data custodians want to make data as available as possible for analysis.
  - Maximises the use of the data…

BUT

- They must protect the confidentiality of respondents.

Systems for flexible output tables can provide protection to the data as well as customization by the analyst

Australian Bureau of Statistics

Australian Bureau of Statistics

2011 Census - Employment, Income and Unpaid Work

STATE, AGE5P - Age in Five Year Groups, Sex (SEXP), Highest Year of School Completed (HSCP) and Indigenous Status (INGP) by Labour Force Status (LFSP)

Counting: Persons, Place of Usual Residence

Filters:

Default Summation Persons, Place of Usual Residence

Labour Force Status (LFSP)						Employed, worked full-time	Employed, worked part-	Employed, away from work	Unemployed, looking for full-	Unemployed, looking for part-	Not in the labour force	Total
	STATE	AGE5P - Age in Five Year Groups	Sex (SEXP)	Highest Year of School Completed (HSCP)	Indigenous Status (INGP)							
				Year 12 or equivalent	Non-Indigenous	103010	21338	5958	4863	1804	13242	150215
					Aboriginal	946	180	88	88	12	184	1498
					Torres Strait	38	5	0	0	0	3	46
					Both Aboriginal	12	0	0	16	0	10	38
					Not stated	579	189	89	59	3	147	1066
				Year 11 or equivalent	Non-Indigenous	8643	1382	586	685	121	1151	12568
					Aboriginal	237	54	29	82	3	75	480
					Torres Strait	13	0	0	0	0	0	13
					Both Aboriginal	0	10	0	0	0	0	10
					Not stated	47	12	12	20	3	16	110
				Year 10 or equivalent	Non-Indigenous	23641	3581	1779	2074	325	3353	34753
					Aboriginal	741	148	66	175	15	331	1476
					Torres Strait	8	3	0	4	11	3	29
					Both Aboriginal	5	3	0	0	0	3	11
					Not stated	170	29	26	29	3	26	283
				Year 9 or equivalent	Non-Indigenous	2529	612	226	706	91	1124	5288
					Aboriginal	161	54	10	111	14	221	571
			Male		Torres Strait	3	0	0	3	0	3	9
					Both Aboriginal	7	0	0	3	0	0	10
					Not stated	30	4	3	3	3	6	49
				Year 8 or below	Non-Indigenous	668	245	88	224	42	755	2022
					Aboriginal	45	14	7	44	3	111	224
					Torres Strait	0	0	0	0	0	0	0
					Both Aboriginal	0	0	0	0	0	0	0
					Not stated	14	3	0	3	3	18	41
				[illegible]	Non-Indigenous	262	114	33	53	7	359	828
					Aboriginal	0	0	0	0	0	9	9

DATA CUBE

- With a little bit of effort we can turn this into a ‘data cube’.

State	Agegrp	Sex	Education	Indigenous Status	LFS Status	Count
NSW	25-29 years	Male	Year 12 or equivalent	Non-Indigenous	Employed	130306
NSW	25-29 years	Male	Year 12 or equivalent	Indigenous	Employed	1269
NSW	25-29 years	Male	Year 12 or equivalent	Not stated	Employed	857

- Collapsed some categories…
- Set-up to model unemployed vs employed ***but can easily handle the three category multinomial outcome***…

DATA CUBE

- This can then be analysed just like a dataset of individual records.

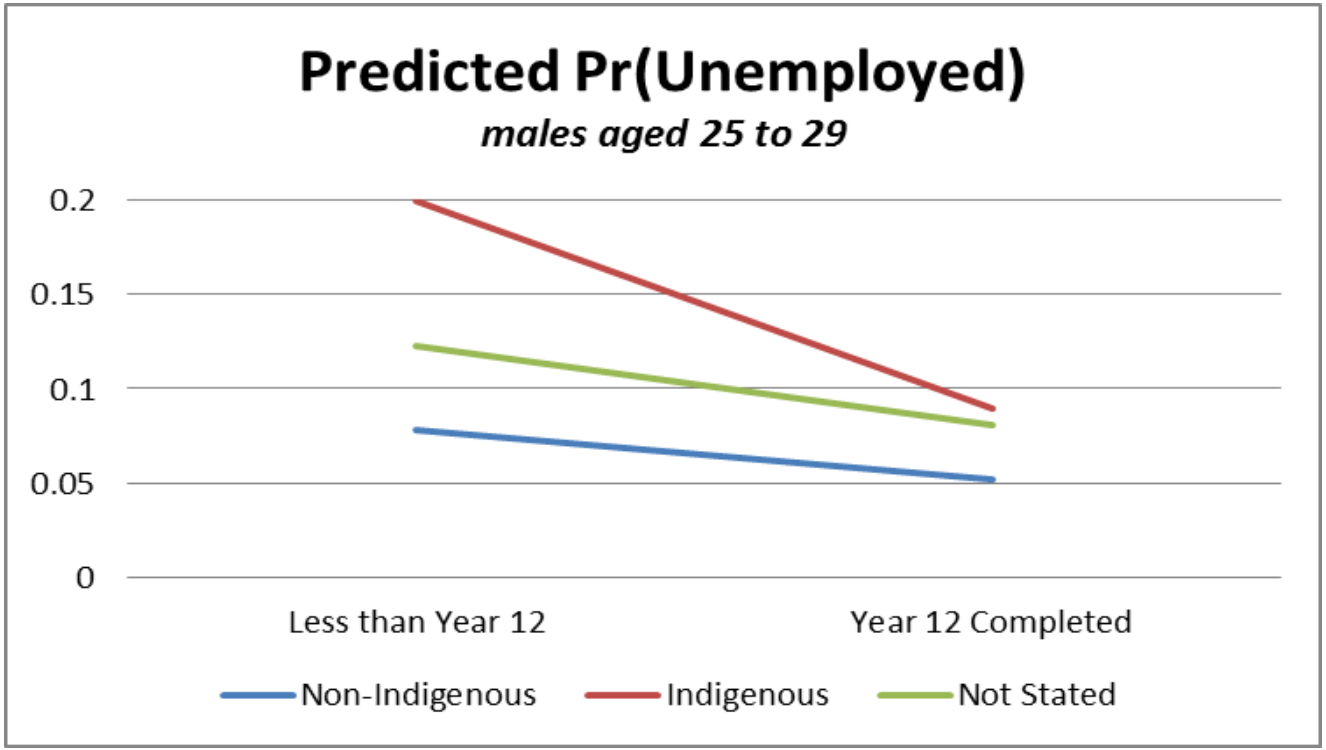
```
Proc logistic data=work.import;
    freq count;
    class agegrp (param=reference ref=first) sex (param=reference ref=last)
          educ (param=reference ref=first) status (param=reference ref='Non-Indigenous');
    model lfs_status (descending) = agegrp sex educ status educ*status;
run;
```

- *In this case with SAS it just ‘repeats’ each combination by the frequency variable.*

MODEL OUTPUT; PREDICTING UNEMPLOYMENT

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Age 30-34 years vs 25-29 years	0.808	0.791	0.824
Age 35-39 years vs 25-29 years	0.733	0.719	0.748
Age 40-44 years vs 25-29 years	0.679	0.665	0.693
Age 45-49 years vs 25-29 years	0.620	0.607	0.633
Age 50-54 years vs 25-29 years	0.579	0.567	0.592
Age 55-59 years vs 25-29 years	0.584	0.570	0.597
Age 60-64 years vs 25-29 years	0.643	0.626	0.660
Female vs Male	1.073	1.061	1.086
Year 12 or equivalent vs Less Than Year 12	0.638	0.630	0.645
Indigenous vs Non-Indigenous	2.627	2.549	2.707
Not stated vs Non-Indigenous	1.634	1.547	1.726

ADDING AN INTERACTION…



DEVELOPING THE IDEA…

- Modern administrative data often comes from different databases.
  - Merge them into a single dataset for analysis.
    - Privacy concerns / Large datasets.

Ideally, we would like to model directly from different databases without having to combine them.

MODELLING APPROACH

- Started with thinking about (relatively) rare events.
  - Heart disease (hospital database, census database, area information)…
  - Unemployment…
- Poisson model a good approximation to logistic in this case.
  - Nice parameter interpretation.
  - Can also then deal with counts.
- Like to be able to deal with between area variability.

GENERALISED LINEAR MIXED MODELS (GLMMS)

y_{ij} : number of events for **observation i in domain j**

$$y_{ij} \mid u_j \sim Pois(\lambda_{ij})$$

$$\lambda_{ij} = N_{ij}\, exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha} + u_j)$$

$$u_j \sim N(0, \sum)$$

- Normal distribution is convenient for interpretation and for the incorporation of correlation structure…
  - **But other distributions are possible…**

‘CONJUGATE’ GENERALISED LINEAR MIXED MODELS (CGLMMS)…

y_{ij} : number of events for **observation i in domain j**

$$y_{ij} \mid u_j \sim Pois(\lambda_{ij})$$

$$\lambda_{ij} = N_{ij}\, u_j\, exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})$$

$$u_j \sim Ga(\mu_j, \kappa\mu_j)$$

Gamma distribution multiplying on original scale.

- Seems a natural way to capture the between area variation.
- Area level covariates enter the area equation to adjust μ_j.

Approach implemented in Lee, Brown, and Ryan (2017).

LOG-LIKELIHOOD – *GAMMA*

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

LOG-LIKELIHOOD – *GAMMA*

Summary statistics of individuals with event

$$\ell(\boldsymbol{\alpha}, \gamma, \kappa; \boldsymbol{y}) \propto \sum_j \left(\sum_{i:Y_{ij}=1} \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} \right)$$

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

LOG-LIKELIHOOD – *GAMMA*

Summary statistics of individuals with event

$$\ell(\boldsymbol{\alpha}, \gamma, \kappa; \boldsymbol{y}) \propto \sum_j \left(\sum_{i:Y_{ij}=1} \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} \right)$$

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

Number of event in area j

LOG-LIKELIHOOD – *GAMMA*

Summary statistics of individuals with event

$$\ell(\boldsymbol{\alpha}, \gamma, \kappa; \boldsymbol{y}) \propto \sum_j \left(\sum_{i:Y_{ij}=1} \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} \right)$$

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

Population at risk

Number of event in area j

Individuals with Event

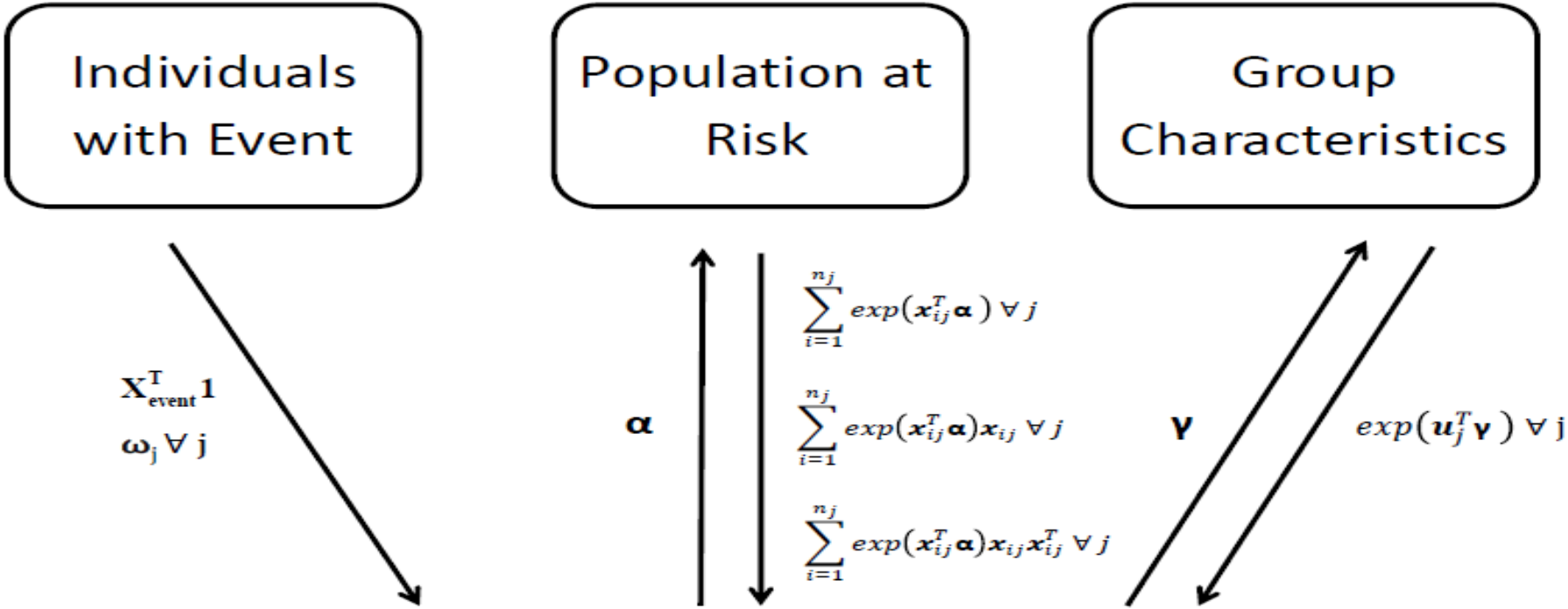
Individuals with Event

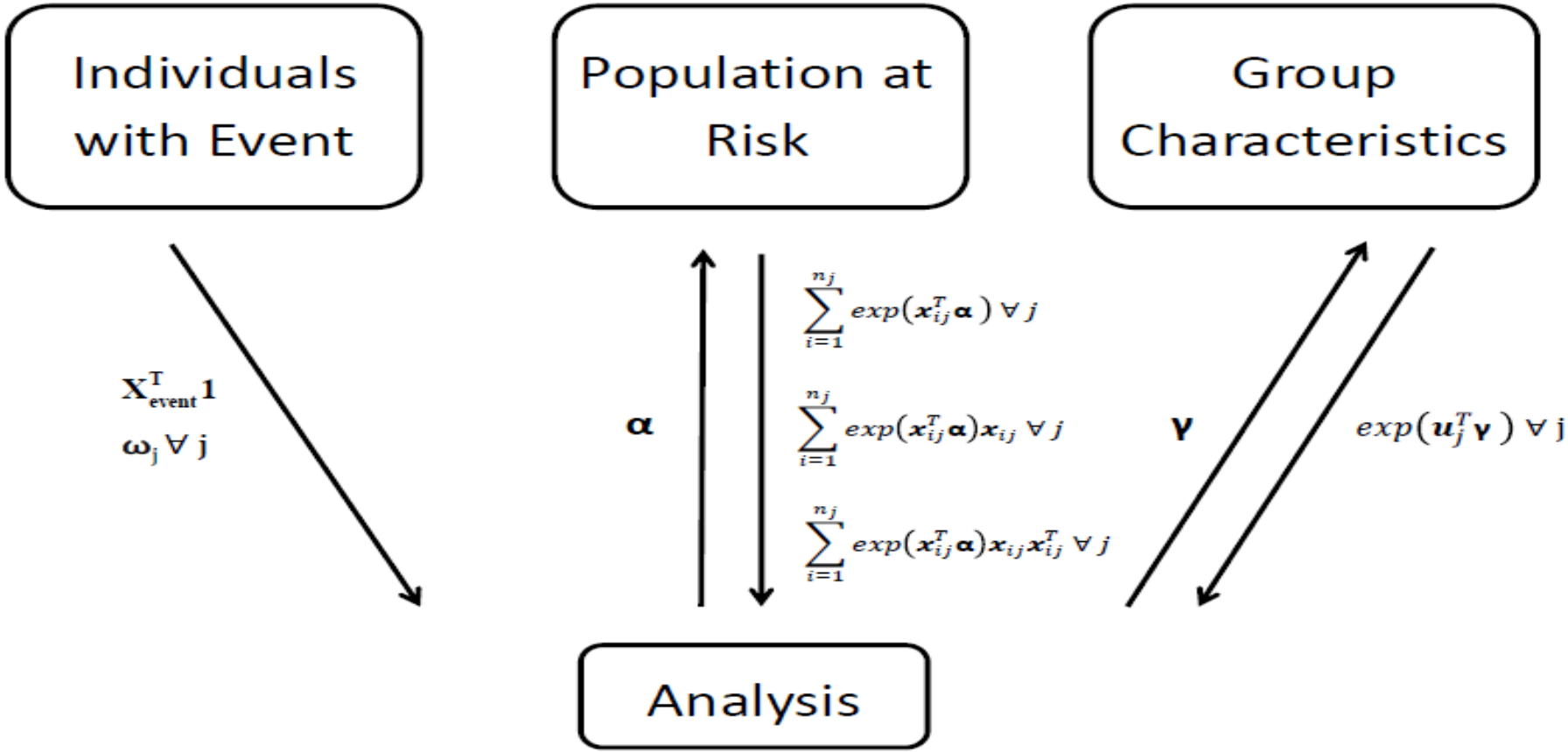
Population at Risk

Individuals with Event

Population at Risk

Group Characteristics


Individuals with Event
Population at Risk
Group Characteristics
$\mathbf{X}^{\mathbf{T}}_{\mathbf{event}}\mathbf{1}$
$\boldsymbol{\omega}_j \forall$ j
$\boldsymbol{\alpha}$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha}) \forall j$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij} \forall j$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T \forall j$
$\boldsymbol{\gamma}$
$exp(\boldsymbol{u}_j^T\boldsymbol{\gamma}) \forall$ j


Individuals with Event
Population at Risk
Group Characteristics
$\mathbf{X}^{\mathbf{T}}_{\mathbf{event}}\mathbf{1}$
$\boldsymbol{\omega}_j \forall$ j
$\boldsymbol{\alpha}$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha}) \forall j$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij} \forall j$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T \forall j$
$\boldsymbol{\gamma}$
$exp(\boldsymbol{u}_j^T\boldsymbol{\gamma}) \forall$ j
Analysis

GLMM

A

Hypothetical Individual Level Data

Area	Sex	Age	HS	Indigenous
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	1
A-CH	1	20-24	0	0
A-CH	1	20-24	0	0
---	---	---	---	---
A-N	1	20-24	1	0
A-N	1	20-24	1	1

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
---	---

GLMM

A

Hypothetical Individual Level Data

Area	Sex	Age	HS	Indigenous
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	1
A-CH	1	20-24	0	0
A-CH	1	20-24	0	0
---	---	---	---	---
A-N	1	20-24	1	0
A-N	1	20-24	1	1

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	HS	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

GLMM

A

Hypothetical Individual Level Data

Area	Sex	Age	HS	Indigenous
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	1
A-CH	1	20-24	0	0
A-CH	1	20-24	0	0
---	---	---	---	---
A-N	1	20-24	1	0
A-N	1	20-24	1	1

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	HS	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

ANALYSIS

CGLMM

B

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
High School Completion (HS)	244479
Indigenous	21900

CGLMM

B

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
High School Completion (HS)	244479
Indigenous	21900

Data on Population at Risk

Area	Sex	Age	HS	Indigenous	N
A-CH	1	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

CGLMM

B

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
High School Completion (HS)	244479
Indigenous	21900

Data on Population at Risk

Area	Sex	Age	HS	Indigenous	N
A-CH	1	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
---	---

CGLMM

B

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
High School Completion (HS)	244479
Indigenous	21900

Data on Population at Risk

Area	Sex	Age	HS	Indigenous	N
A-CH	1	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
---	---

ANALYSIS

GLMM

CGLMM

A

Hypothetical Individual Level Data

Area	Sex	Age	HS	Indigenous
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	1
A-CH	1	20-24	0	0
A-CH	1	20-24	0	0
---	---	---	---	---
A-N	1	20-24	1	0
A-N	1	20-24	1	1

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	HS	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

ANALYSIS

B

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
High School Completion (HS)	244479
Indigenous	21900

Data on Population at Risk

Area	Sex	Age	HS	Indigenous	N
A-CH	1	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
---	---

ANALYSIS

CENSUS UNEMPLOYMENT EXAMPLE

- Used aggregate data from Table-Builder.
  - Modelled at SA4 level (all Australia) to compare GLMM with CGLMM.

  BUT

  - CGLMM needs less detailed data so could work at lower levels (SA3?) without the cell rounding becoming too much of an issue…

- Modelled 2011 unemployment with some individual covariates and a historical area covariate.
  - Individuals with events from 2011 Census.
  - Population at risk from 2011 Census.
  - Area covariate from 2006 Census.

Parameter	Poisson Mixed Model		Gamma-Poisson	
	Est	SE	Est	SE
α_o (Intercept)	-2.06*	0.028	-2.03*	0.027
α_1 (Female)	*Reference Group*			
α_1 (Male)	-0.06*	0.003	-0.06*	0.003
α_2 (age 20 to 24)	*Reference Group*			
α_2 (age 25 to 29)	-0.51*	0.005	-0.51*	0.005
α_3 (age 30 to 34)	-0.70*	0.005	-0.70*	0.005
α_4 (age 35 to 39)	-0.80*	0.005	-0.80*	0.005
α_5 (age 40 to 44)	-0.92*	0.006	-0.92*	0.006
α_6 (age 45 to 49)	-1.03*	0.006	-1.03*	0.006
α_7 (age 50 to 54)	-1.11*	0.006	-1.11*	0.006
α_8 (age 55 to 59)	-1.10*	0.007	-1.10*	0.007
α_9 (Not completed High School)	*Reference Group*			
α_9 (Completed High School)	-0.46*	0.003	-0.46*	0.003
α_{10} (Not Indigenous)	*Reference Group*			
α_{10} (Indigenous)	0.77*	0.007	0.77*	0.007
γ (IRSAD)	-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

CENSUS UNEMPLOYMENT EXAMPLE…

- Unemployment rate reduces for males.
- Unemployment rate reduces by age.
  - *Those NILF are excluded from the analysis…*
- Higher education reduces unemployment rate.
- Strong increase in unemployment for indigenous.
- Residual between area variation.
  - *But historical component of SEIFA not significant…*
  - *State and/or other geographic classification easily added…*

CONCLUDING REMARKS

- Utilising tabular data is nothing new BUT we tend to focus on ‘individual’ record files for modelling:
  - that are often not necessarily needed.
  - that are difficult to handle (size).
  - that have confidentiality issues.

- Using the Gamma model allows us to more easily extract information from different aggregate databases.
  - The likelihood also has a closed-form although computational algorithms still not straightforward…

THANK YOU FOR YOUR ATTENTION