Introduction to Generalized Linear Models: Ordinal Logistic Regression and Poisson Regression

Dr Cameron Hurst
Cameron.Hurst@qut.edu.au

Khon Kaen University

8th September, 2557

Preamble

- We will now move on from the most widely used Generalized Linear Model (Binary Logistic regression), to those used a little less, nonetheless these methods are very important
- You will see that once you understand Binary Logistic regression, the workings of these other Generalized Linear models is very similar
- Although we are now talking about specific models. I would like you to remember the **Unifying theory** that ties them all togther: The Generalized Linear Model

What we will cover....

1. Introduction
2. Modelling ordinal outcomes
3. Poisson regression: Modelling counts and rates

Modelling ordinal and count data

- In this session we will cover methods that conisder a little more information in the outcome variable (i.e. On at least an ordinal measurement scale)
- First, we will cover a method for Ordinal logistic regression: used to model ordinal outcomes.
- Second, Poisson regression which is used for modelling events (especially rare events). These can be in the form of counts (number of events), or rates (e.g. Incidence or mortality rates)

Ordinal Logistic Regression

- Has advantage (over multinomial [nominal] logistic regression) of taking order in an outcome variable into account (when it is present)
  - E.g. Grade of Cancer (progression/severity): Grd I; Grd II; Grd III; Grd IV
- Should be noted that nominal logistic regression is still valid on ordinal outcomes, it is just unlikely to perform as well as the ordinal model **Why???**
- We will consider one of several possible ordinal logistic regression models:

'Proportional odds' ordinal logistic regression

The proportion odds ordinal logistic regression model

- The proportional odds model assumes differences can be represented using the constant term alone (i.e. the other explanatory variables do not depend the categories under consideration). That is:

$$ln\left[\frac{\pi_1+\pi_2+\cdots+\pi_j}{\pi_{j+1}+\pi_{j+2}+\cdots+\pi_J}\right]=\beta_0+\beta_1 X_{i,1}+\beta_2 X_{i,2}+\ldots+\beta_{k-1}X_{i,k-1}+\epsilon_i$$

- That is, the Xs effect the difference between categories in exactly the same way (on a log scale): the effects are proportional.
- Proportional odds model is not the simplest ordinal logistic regression model but it does seem to be the default approach used by most software. Personally, I find it the most intuative

Proportionality

- The idea is: the odds ratio for two adjacent categories (odds of grade I tumour relative to odds of grade II) is effected by the covariates/factors in exactly the same way as OR(grade II/grade III) and so on.
  - E.g. Effect of new drug treatment is the same regardless of which two adjacent categories being considered in OR ($\leftarrow$ Drug always helps).
- This assumption of proportionality is something we will discuss in more detail when we cover the survival analysis model, Proportional Hazards (Cox) regression

Ordinal Logistic Regression:Example

Motivating example - Factors influencing likelihood of postgraduate education

- 400 first year undergraduate students were asked their likelihood of applying for post graduate training: **Apply: Unlikely, Somewhat likely, Very likely** - an ordinal outcome variable.
- Also recorded were:
    1. the students' parental postgraduate training [no, yes],
    2. whether the university the students currently attended were research intensive[no, yes], and
    3. the students GPA [numerical]

Results

Motivating example - Factors influencing likelihood of postgraduate education

apply

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unlikely	220	55.0	55.0	55.0
	Somewhat likely	140	35.0	35.0	90.0
	Very likely	40	10.0	10.0	100.0
	Total	400	100.0	100.0	

ParentPostGrad

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No postgrad	337	84.2	84.2	84.2
	Postgrad	63	15.8	15.8	100.0
	Total	400	100.0	100.0	

LowResearch

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low reserach output	343	85.8	85.8	85.8
	High reserach output	57	14.2	14.2	100.0
	Total	400	100.0	100.0	

Results

Motivating example - Factors influence likelihood of postgraduate education

And cross-tabulations:

ParentPostGrad * apply Crosstabulation

			apply			
			Unlikely	Somewhat likely	Very likely	Total
ParentPostGrad	No postgrad	Count	200	110	27	337
		Expected Count	185.4	118.0	33.7	337.0
	Postgrad	Count	20	30	13	63
		Expected Count	34.6	22.0	6.3	63.0
	Total	Count	220	140	40	400
		Expected Count	220.0	140.0	40.0	400.0

LowResearch * apply Crosstabulation

			apply			
			Unlikely	Somewhat likely	Very likely	Total
LowResearch	Low reserach output	Count	189	124	30	343
		Expected Count	188.6	120.0	34.3	343.0
	High reserach output	Count	31	16	10	57
		Expected Count	31.4	20.0	5.7	57.0
	Total	Count	220	140	40	400
		Expected Count	220.0	140.0	40.0	400.0

Watch out for low frequency cells

Ordinal logistic regression: Result

Motivating example - Factors influence likelihood of postgraduate education

Table 1: The overall model is significant ($p<0.05$)

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	557.272			
Final	533.091	24.180	3	.000

Link function: Logit.

Table 2: The model predictions do no deviate significantly from the data (the model provides a good fit)

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	400.843	435	.878
Deviance	400.749	435	.879

Link function: Logit.

Ordinal logistic regression: Result

Motivating example - Factors influence likelihood of postgraduate education

Table 3: The Nagelkerke R^2 is not very high (but how high should it be ????)

Pseudo R-Square

Cox and Snell	.059
Nagelkerke	.070
McFadden	.033

Link function: Logit.

Now for the Pointy End: Coeffcients interpretation

Interpreting βs and ORs from Ordinal Logistic models

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[apply = .00]	2.203	.784	7.890	1	.005	.666	3.741
	[apply = 1.00]	4.299	.809	28.224	1	.000	2.713	5.885
Location	gpa1_4	.616	.263	5.499	1	.019	.101	1.130
	ParentPostGrad	1.048	.268	15.231	1	.000	.522	1.574
	LowResearch	-.059	.289	.041	1	.839	-.624	.507

Link function: Logit.

- First, just ignore the threshold: these are just group specific interecpts (i.e.β_0s)
- Second, we can only see the coeffcients (βs) not the odds ratios, but:
  - Both **GPA** and **Parents postgrad training** are associated with the students likelihood of applying for postgrad training (both $p < 0.05$)
  - However, the **research intensity of the university** was not.

Interpreting βs and ORs from Ordinal Logistic models

To get the ORs (and their 95%CIs) we need to expontiate the βs (and their 95%CIs). For GPA:

$$OR_{GPA} = e^{\beta_{GPA}} = exp(0.616) = 1.852$$

and the 95%CI:

$$]e^{0.101}, e^{1.13}] = [1.11, 3.1]$$

So the odss of being one category higher of 'likelihhod for future application for posgraduate training' increases by 85 % for every extra additional GPA point.

We can do the same for all other coeffcients and their 95%CIs, but I won't bother.

Performing Ordinal Logistic regression in R

Very similar to conducting a standard logistic regression, except we use the `polr` function in the `MASS` R library:

Proportion odds logistic regression in R

```
library(MASS)

my.ordlr<-polr(my.y~my.x1+my.x2, data=mydata.df)
summary(my.ordlr)
```

Ordinal logistic regression

- Ordinal logistic regression might seem a bit tricky at first, but once you get used to them, they are pretty straight forward
- If you don't do them that often, you will find you forget how they work
- My suggestion is that the first time you run one, spend a little time 'writing it up', then for the next one you do, refer back to your write up.

Poisson regression

- Personally, I find Poisson regression somewhat more accessible than the logistic regression methods (then why didn?t I cover it first?)
- However, the use of this method is (surprisingly) not as common in health and clinical research (especially as binary logistic regression)
- This is probably because researchers tend to prefer dichotomous outcomes (even when it makes more sense to model counts)

Poisson regression: What I will cover

- Counts, rates and the Poisson distribution
- When the (standard) Poisson distribution doesn?t fit
- Counts vs rates: the idea of an offset.

The Poisson distribution

- Health outcomes that are counts (especially rare events) often have a Poisson distribution
  - Number of road accidents per day
  - Number of lesions
- These counts have to be represented by natural numbers (i.e. Positive integers): 0, 1, 2,...
- As the natural numbers are zero bound (negative values not possible), and higher numbers are possible, but unlikely, the Poisson distribution is often positively skewed

Poisson vs. the Normal distribution

- The normal distribution has a parameter that represents it?s centre (μ, the Mean) and one that represents it?s spread (σ^2, the Variance or σ, the Standard deviation)
- In contrast, the Poisson has a single parameter (λ, Lambda) that represents both the mean and the variance
- In other words level of mean of a Poisson distribution is the same as the level of the variance
- IN FACT, we can use this property to identify the Poisson distribution.

When do we use Poisoon regression?

ANS: When we have counts of rare events (counts)

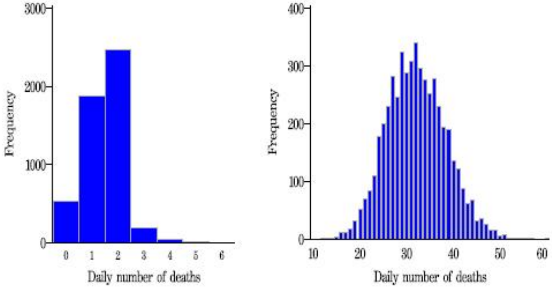
- BUT if counts are not so rare, we can often use general linear model (i.e. Linear regression instead)
- General linear models (linear regression) avoids the complexity introduced by (log) link functions and Maxmimum likelihood estimation
- Generally, if the mean count (or $\lambda = \mu$) > 15, we can use standard linear regression (check data using histogram)

Note: Linear regression and Poisson regression

When the count outcome is 'not so rare' (e.g. $\lambda > 15$), when can use Linear regression instead of Poisson regression

Rare vs common events

To illustrate this "$\mu > 15$" rule, let's consider the daily number of deaths (from a particular disease) in Bangkok and Khon Kaen. BKK is large and has 32.1 deaths per day, on average (i.e. $\lambda = 32.1$), whereas KK which is much smaller has (on average) 1.2 deaths per day (i.e. $lambda = 1, 2$)



Number of deaths for BKK appears normally distributed and we note: $\lambda = 32.1 > 15$

Poisson regression as a Generalized Linear Model

The **link function** most used for the Possion regression, that is the **Canonical or Natural link**, is the **Log** function.

Fortunetly, unlike the Logit link used for Logistic regression, the natural log function, is a little easier to undersatnd.

Let's just quickly reconsider the definition of the Generalized Linear model...

The Generalized Linear Model

DEFINITION: A *Generalized Linear Model* is a model that can be represented:

$$g(y) = \mathbf{X}\beta + \epsilon$$

Or equivalently,

$$y = g^{-1}(\mathbf{X}\beta + \epsilon)$$

where $g()$ **monotonic** and **differentiable** link function

For **Poisson regression**, y is often a count variable, so:

$$g(y) = ln(y) = \mathbf{X}\beta + \epsilon$$

or alternatively,

$$y = e^{\mathbf{X}\beta + \epsilon}$$

R syntax for Poisson regression

You should be able to guess!!!!!

Poisson regression in R

```
my.poireg<-glm(my.y~my.x1+my.x2, data=mydata.df,
family=poisson)
summary(my.poireg)

# Use print.RRCIsi to get RRs and their CIs
print.RRCIs(my.poireg)
```

THAT'S RIGHT!!!! EXACTLY THE SAME AS A BINARY LOGISTIC REGRESSION EXCEPT `family=poisson` **RATHER THAN** `family=binomial`

Example: Death rate and temperature

We will considers the rates of deaths in different cities. We believe that the minimum winter temperatures can help explain mortality

QUESTION: Why are we considering rates (and not just raw counts????)

Results

In this case, we can see that there is a significant temperature effect ($RR = 0.987$, $p < 0.001$, 95%CI: 0.984, 0.989). As temprature increases 1 degree, there is a (1-0.987)x100% =1.3% decrease in the rate of deaths

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)
			Lower	Upper	Wald Chi-Square	df	Sig.	
(Intercept)	3.536	.0067	3.522	3.549	274427.627	1	.000	34.312
tmpd10	-.013	.0013	-.016	-.011	111.480	1	.000	.987
(Scale)	1[a]							

Dependent Variable: death
Model: (Intercept), tmpd10

a. Fixed at the displayed value.

Parameter Estimates

Parameter	95% Wald Confidence Interval for Exp(B)	
	Lower	Upper
(Intercept)	33.862	34.769
tmpd10	.984	.989
(Scale)		

Dependent Variable: death
Model: (Intercept), tmpd10

Overdispersion and Underdispersion

Overdispersion occurs when there is too much variance (for the Poisson distribution).

Recall for a Poisson distribution, there is a single parameter, λ, and

$$\lambda = \mu = \sigma^2$$

Overdispersion is when

$$\sigma^2 > \mu$$

Underdispersion (rarer, but still possible) is when

$$\sigma^2 < \mu$$

Dealing with Over- and Under-dispersion

There are a couple of methods of dealing with Over- and Under-dispersion. Perhaps the best approach is to use the:

Negative Binomial Generalized Linear Model

This is a two parameter Generalized Linear Model (has both a location and scale parameter). One of the problems is to estimate what the dispersion (scale) parameter is first, and then fix this in the model.

I won't bother going through this model. If you get this problem, hit the books (or come to see me)

Offsets

If we consider two cities, Bangkok and Khon Kaen in our analysis. We want to model the number of mortalities in both these cities and see if there is a difference between the two cities.

You may recall that we observed about **32.1 deaths per day in BKK**, and **1.2 deaths per day in KK**. Does this mean that the **risk** of mortaility (from our disease) is lower in KK?

Using an offset

- We would expect, Khon Kaen to have a much lower number of deaths (population = 200 000), simply because it is a smaller city (even though the mortality rate might be higher in KK).
- In order to deal with this we can include an offset variable in the model.
- For example, a sensible offset might be represented by the city populations

Poisson regression models with and without an offset

Where we don't include an offset our model is:

$$ln(count) = \mathbf{X}\beta + \epsilon$$

To include an offset:

$$ln(count/population) = \mathbf{X}\beta + \epsilon$$

In the top model, we are modelling **counts** (number of events), in the bottom model (with an offset) we are model **rates** (Incidence rate, Mortaility rate etc.)

Concluding remarks

- Poisson regression is the best method to use for outcomes represented by (rare) counts
- If the mean number of counts is large ($\lambda = \mu > 15$) then you might be able to use standard general linear model instead (can be simpler)
- Use the Negative binomial GLM to account for response data that doesn?t perfectly fit the Poisson restriction that the mean is equal to the variance
- Use the offset to account for different exposure sizes

THANK-YOU

Questions?????