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

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

24th August, 2558

What we will cover....

Outcomes and Scales of measurement

We have now covered models for the analysis of **Continuous** outcome (Linear regression, ANOVA and the general linear model)

We have also covered a method for the analysis of data at the other end on the measurement scale spectrum, **Binary** outcomes, using Binary logistic regression

What about if we have data somewhere between continuous data (fully quantitative), and the more qualitative binary outcome. For example:

- We might have a **Nominal** (NAME) variable (e.g. Type of cancer)
- An **Ordinal** outcome: Grade of cancer (I, II, III, IV and V)
- Or something more quantitative like a (rare) **Count** (e.g. Number of cases of malaria in an amphour). As a rare count it is very unlikely to be normally distributed

Modelling nominal, ordinal and count outcomes

- In this session we will cover methods that provide a little more information (i.e. On at least an ordinal measurement scale)
- We will first consider methods for the nominal outcome case.
- Second, we will cover a method for Ordinal logistic regression: used to model ordinal outcomes.
- Third, Poisson regression, used for modelling events (especially rare events). These data are represented by counts (number of events), or rates (e.g. Incidence rate or mortality rate)

Take home message(Again)

The nature (scale and distribution) of the outcome is THE most important factor in choosing the right model. UNDERSTAND THIS AND YOU ARE MOST OF THE WAY TO UNDERSTANDING BIOSTATISTICAL MODELING

A quick review of the binary lositic regression

Recall where we have a binary outcome variable which can take values 0 and 1, we would employ a binary logistic regression

$$\log\left(\frac{P(Y=1|x)}{P(Y=0|x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1}$$

or equivalently,

$$\frac{P(Y=1|x)}{P(Y=0|x)} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_{k-1} X_{k-1}}$$

Simplifying (matrix notation):

$$\frac{P(Y=1|x)}{P(Y=0|x)} = e^{X\beta}$$

We would then fit this glm model to get estimates of the coeffcients, $\hat{\beta}$

Review of binary logistic regression

- In a **binary** logistic regression, we are usually considering the odds of an event (e.g. occurance of a disease) against a non event (disease-free)
- We consider these two odds (odds of disease vs odds of non-disease) in a ratio: the odds ratio
- We are most interested this in terms of a risk factor (or treatment)
- We then use a Generalized linear model to estimate the odds ratio(s) and their confidence interval
  - Remember: A CI that does not contain 1 suggests that a risk factor is significantly associated with the odds of a disease

Modelling nominal outcomes

- What about if our outcome is on a nominal scale: Three or more non-ordered categories
- For example, we might consider 'type' of cancer: Liver cancer; Stomach cancer; and Colo-rectal cancer
- In this case we would need to employ a method that can deal with the multiple (unordered) categories of the outcome variable: **Multinomial (Nominal) Logistic regression**

Multinomial (Nominal) Logistic regression outcomes

In a way similar to how we consider a referent category for risk factors, we choose a particular outcome in our nominal outcome variable as the 'baseline' comparator

For example we might consider 'Liver cancer' as the baseline comparator, against which we compare the odds of the other two types of cancer (Stomach cancer; and Colo-rectal cancer)

In this way, we can consider a Nominal logistic regression as a **set of binary logistic regressions**

Nominal logistic regression model

Now, let's say we have a nominal outcome with g categories (e.g. $g = 3$ for cancer type: Liver, Stomach, Colo-rectal). We will then need $g - 1$ models. Each model considers the odds of a particular outcome, against a 'baseline' outcome:

$$\log\left(\frac{P(Y=1|x)}{P(Y=G|x)}\right) = e^{X\beta_1}$$

$$\vdots$$

$$\log\left(\frac{P(Y=G-1|x)}{P(Y=G|x)}\right) = e^{X\beta_{G-1}}$$

β_1 represents the log odds for 'outcome 1' relative to 'outcome G', and β_{G-1} the log odds of 'outcome G-1' relative to 'outcome G'. **We can see this as simply a set of binary logistic regression models.**

Nominal logistic regression model: Example

So for our cancer example: Liver Cancer(LC) , Stomach Cancer(SC), and Colo-rectal cancer(CRC), and using colo-rectal cancer as the 'baseline' cancer: We will have $g - 1 = 3 - 1 = 2$ models.

$$\log\left(\frac{P(Y = LC|x)}{P(Y = CRC|x)}\right) = e^{X\beta_{LC}}$$
$$\log\left(\frac{P(Y = SC|x)}{P(Y = CRC|x)}\right) = e^{X\beta_{SC}}$$

β_{LC} represents the log(ORs) for LC relative to group CRC; and β_{SC} the log(ORs) odds of SC relative to CRC.

Multinomial logistic regression in R

Many of the logistic regression models in R are a pain (they expect to be dummy coded using g columns), I prefer to have a single variable representing the nominal outcome (Letting R work out the dummy coding).

The R library `nnet` provides a multinominal logistic regression routine that will accept a single nominally coded variable for the output:..

`nnet` comes with R, but we still have to load it into the memory using the library statement: `library(nnet)`

Hypertensive status and achievement of particular targets

You should recall from our DMHT dataset, that three clinical targets are being considered:

A `a1cyn`$\rightarrow$ Hemoglobin A1c (yes: $< 7\%$; no $\geq 7\%$)

B `bpyn`$\rightarrow$ Bloop pressure(yes: $< \frac{130}{80}$mmHG; no $\geq \frac{130}{80}$ mmHg)

C `ldlcyn`$\rightarrow$ Low density lipoprotien-Cholesterol(yes: $<$ 100mg/dL; no $\geq$ 100mg/dL)

For those that achieve only one clinical target, does the HT status of the patient matter (in terms of WHICH clinical target the achieve)? Also, does diabetes duration relate to the idenitity of this single clinical target?

Case study: Research question

So to clarify, where have two main research questions:

1. Does hyptertention (status) relate to which clinical target (A, B or C) is achieved
2. Does time since diagnosis (of diabetes) relate o which clinical target (A, B or C) is achieved

After generating this variable and dataset (sample representing those achieving a **single** clinical target), n=689.

Note: I have provided a script file of both the data preparation and analysis in: `NomOrdPoissonReg.R`

Case study: Results

R code: Multinomial (Nominal) logistic regression

```
mn.reg <- multinom(ABC.fac ~ ht + duradm, data
= dmht.oneTarget.df)
summary(mn.reg)
# User-defined function (ORs, CIs, p-vales)
printout.mnlogreg(mn.reg)
```

The function `printout.mnlogreg` is a function I wrote to generate the ORs, 95%CIs and Wald test p-values. See the end of the lecture about how to access this function.

Case study: Results

Hint: Just consider the output in terms of two binary logistic regressions (that happen to share the same 'baseline' category).

Case 1: HbA1c vs BP)

HbA1c (vs BP)	OR	95CI	p (Wald test)
Hypetertension	3.17***	(2.04, 4.93)	< 0.001
DM duration	0.92***	(0.87, 0.97)	< 0.001

- We see that the odds of Blood sugar control (compared to BP achievement) is 3.17 times higher in Hypertensive patients, than in non hypertensive patients
- T2DM duration seems to reduce the odds of achieving blood sugar control (vs BP) (In other words, as time progresses the patients ability to control blood sugar, compared to control of BP, decreases)

Case study: Results

Case 2: Cholesterol vs BP)

LDLC (vs BP)	OR	95CI	p (Wald test)
Hypetertension	2.49***	(1.70, 3.62)	< 0.001
DM duration	1.01	(0.97, 1.04)	0.29

- The odds of cholesterol target achievement (compared to BP achievement) is 2.5 times higher in Hypertensive patients, than in non hypertensive patients (as we would expect, HT patients find it easier to achieve cholesterol control, than BP control)
- T2DM duration seems to have no effect on the chance of achieving chlesterol (vs BP)

A last comment on Nominal logistic regression

- Personally, I find multinomil logistic regression cumbersome
- The use of multiple models (within a single analysis) tends to lead to convoluted explanations
- The trick is to think of the 'sub' models as separate binary logistic regression models (and interpret then correspondingly)

Take home message: Nominal logistic regression

Think of multinomial (nominal) logistic regression as a 'set' of binary logistic regression analyses, and interpret each model seperately.

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 ordinal **WHY?????**
- We will consider one of several possible ordinal logistic regression models:

'Proportional odds' Logistic regression

The proportionality assumption

- 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). The model is:

$$ln\left[\frac{\pi_1+\pi_2+\cdots+\pi_j}{\pi_{j+1}+\pi_{j+2}+\cdots+\pi_J}\right]=\beta_0+\beta_1X_{i,1}+\beta_2X_{i,2}+\ldots+\beta_{k-1}X_{i,k-1}+\epsilon_i$$

- Specifically, the Xs effect the difference between categories in exactly the same way (on a log scale). In other words, 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

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

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 results: Coeffcients

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.

From the table (wich give βs not ORs), we can see:

- The odds of (intension) of postgraduate study appears to goes up with GPA ($\beta = 0.616$; $OR = e^{0.616} = 1.8$, $p = 0.019$)
- Parental postgraduate training also made intention for postgraduate traing more likely ($p < 0.001$)
- BUT there was no evidence that attendance at a research intensive (high ranking) university influence intension of PG training

Interpretation of proportional odds logistic regression ORs

- Compared to nominal logistic regression models, proportional odds logistic regression simple to interpret.
- We can just say **" The odds of 'intension for post graduate training being ONE LEVEL higher for risk factor X is...."**
- In every other respect, they are like Binary logistic regression
- It is important (in our selection of the proportional odds method) that the assumption of proportionality is valid
- We can do this simply by running binary logistic regression on two subsamples (The fisrt is 'somewhat likely' vs 'unlikely", the second being 'very likely' vs 'somewhat likely'). If the βs (or ORs) are similar, than proportionailty assumption is safe.
- IF IT ISN'T, we need to consider another type of ordinal logistic regression model, or (IF WE HAVE TO) go to a multinomial logistic regression

R code for Proportional odds logistic regression

R syntax: Proportional odds logistic regession

```
library(MASS)
my.plr <- polr(my.y~my.X1+my.X2, data =
mydat.df)
summary(my.plr)
```

Two things to note:

1. The proportional odds model in R is part of the MASS library (comes with R, but still has to be loaded)
2. This syntax only gives the βs. We would have to calculate the ORs and their CIs (maybe I will write a function soon)

Rare counts and rates

- Often we conduct ecological studies where we have aggregated data (e.g. Counts of events at the hospital or province level)
- For example, we might consider counts of cases of liver cancer at the Amphour level
- As liver cancer is comparatively rare, especially in samples not that large (a few thousand people), they are unlikely to be normally distributed

Rule of thumb: Poisson vs Normal distributions

As a general rule, when the average count is < 15 we cannot assume normality, and we should consider a Poisson (or Negative binomial) distribution

Counts vs (incidence) rate

It may also not be that sensible to consider counts (because of different sized sampling units), instead we may want to consider incidence or mortality rates

For example:
The count of liver cancer cases (at a given time) in BKK might be 1500, whereas in Khon Kaen it might be 23. Does this mean it is much riskier to live in BKK (in terms of liver cancer)?

To offset the difference in the number of susceptables, we might consider incidence rate per 10,000 person-years. Now the incidence in BKK is 15 per 10,000 person-years, whereas in KK it is 16.5 per 10,000 person-years.

Poisson regression can be used in either case

The Possion regression model

There are two possible models:

1. In the first model, we don't have to offset sampling unit size:

$$log(y) = X\beta$$
$$\Rightarrow y = e^{X\beta}$$

The Possion regression model

2. Or, if sampling units have different numbers of individuals (e.g. Comparing different amphour: KK vs BKK)

$$\log(\frac{y}{n}) = X\beta$$
$$\Rightarrow \frac{\log(y)}{\log(n)}) = X\beta$$
$$\Rightarrow \log(y) - \log(n) = X\beta$$
$$\Rightarrow \log(y/n) = \log(n) + X\beta$$

In either case (as with Binary logistic regression), we simply expontiate our βs to get our RATE ratios:

$$RR = e^{\beta}$$

The effect of Hypertension on number of targets achieved (A,B,C)

Using the DMHT2500 data, I will remove any observations which are missing for A, B or C:

R suntax: Create new dataset

```
#CREATE VARIABLE WHICH IDs COMPLETE CASES
cc.ABC<-complete.cases(dmht.df$a1cyn,
dmht.df$bpyn, dmht.df$ldlcyn)

#BUILD A NEW DATASET WITH ONLY COMPLETE CASES
#dmht.ccABC.df<-dmht.df[cc.ABC,]
```

R code for Poisson regression

Poisson regression is a generalized linear model, and as such, we can specify it using R's native `glm` function

R suntax: Poisson regression (no offset)

```
my.pois.reg<-glm(numABC~ht+duradm,
data=dmht.ccABC.df, family=poisson)
summary(my.pois.reg)

#Calculate rate ratios and their 95% CIs
#using the user-defined function...
print.RRCIs(my.pois.reg)
```

Note all of these user defined functions can be found in `DMHT2500v2.Rdata`, or run using the source code file `Rutilityfunctions.R` which I have put in dropbox

```
> my.pois.reg<-glm(numABC~ht+duradm, data=dmht.ccABC.df, family=poisso
> summary(my.pois.reg)

Call:
glm(formula = numABC ~ ht + duradm, family = poisson, data = dmht.ccAB

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.5806  -0.2296  -0.1597   0.6915   1.4396

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.225110   0.052573   4.282 1.85e-05 ***
ht          -0.061615   0.049047  -1.256    0.209
duradm      -0.001316   0.004759  -0.276    0.782
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1307.2  on 1577  degrees of freedom
Residual deviance: 1305.5  on 1575  degrees of freedom
(50 observations deleted due to missingness)
AIC: 4119.3
```

```
Number of Fisher Scoring iterations: 5

> print.RRCIs(my.pois.reg) #User-defined function
RR    RR.L95   RR.U95
(Intercept) 1.2524600 1.1298283 1.388402
ht          0.9402451 0.8540662 1.035120
duradm      0.9986853 0.9894127 1.008045
```

Case study: Results

- From the **coeffcients table** R output we can see neither Hypertension or T2DM duration could be shown to associated with the number of targets achieve (0.1,2,3).
- Going to the RRs we can see the rate of 'clinical target achievement' is 6%[$100\%x(1-0.9402)$] less for hypertensive patients, compare to non hypertensive DM pateinst **BUT NOT SIGNIFICANT** (i.e. I am only interpreting for an example)

Any problems????

- What type of distribution do you think the variable, 'Number of clinical targets achieved' has
- Like all count variables it is bound 'on the left' by zero (you can't acheive 'negative numbers of achievements')
- BUT unlike many count variables, it is also bound 'on the right'. The maximum number of achievements is 3
- That is, *Number of clinical targets achieved* can only take values: 0, 1, 2, 3
- This raises doubts about the validity of the Poisson model.

Dealing with problems with the Poisson distribution

- This boundary on both ends of our outcome variable is likely to lead to a particular type of problem: Over- or Under-dispersion
- One of the properties of the Poisson distribution is that the mean MUST equal the variance
- That is, a single parameter, λ, can be used to specify the Poisson distribution because: $\lambda = \mu = \sigma^2$
- What about if the variance is too high: $\sigma > \mu$ (Overdispersion)
- or the variance is too low: $\sigma < \mu$ (Underdispersion)
- We can 'fiddle' with the parameters and fit an over (or under) dispersed Poisson regression;
- OR we could use a different distribution (that has two parameters) which may fit the data better: Such as a **Negative binomial distribution**

Negative binomial regression in R

I am not going to bother going through a full analysis using an overdispersed Poisson regression, or a Negative binomial regression. BUT I will provide you the code, just in case you ever want to run one:

R suntax: Negative binomial regression

```
library(MASS) #contains glm.nb function

#glm.nb specific to neg bin, so no family
my.nb.reg<-glm.nb(numABC~ht+duradm,
data=dmht.ccABC.df)
summary(my.nb.reg)
```

BTW: I do run the negative binomial model in the syntax file associated with this lecture: `NomOrdPoissonReg.R`

User-defined functions, Data, R script for this lecture

I have provided you access to functions I have specially written (e.g. `Print.ORCIs`). You can access this two ways:

1. The source code (a file containing about 20-30 functions that just needs to be run prior to using the functions).
2. Or just ope the syntax file in R and run the code.
   - You can do this by including the line `source('RutilityFunctions.R')` at the start of your R code (assuming you have put this file into your working directory)
3. Or you can just open the workspace `DMHT2500_Version2.Rdata` Remember you can do this two ways:
   1. Including the line `load('DMHT2500_Version2.RData')` in your code
   2. Opening this RData file into the 'environment' window in RStudio (top-right pane)

THANK-YOU

Questions?????