Modelling Rates

Mark Lunt

Centre for Epidemiology Versus Arthritis
University of Manchester

06/12/2022

Modelling Rates

- Can model prevalence (proportion) with logistic regression
- Cannot model incidence in this way
- Need to allow for time at risk (exposure)
- Exposure often measured in person-years
- Model a rate (incidents per unit time)

Assumptions

- There is a rate at which events occur
- This rate may depend on covariates
- Rate must be ≥ 0
- Expected number of events = rate $\times$ exposure
- Events are independent
- Then the number of events observed will follow a Poisson distribution

Poisson Regression

- Negative numbers of events are meaningless
- Model $\log(rate)$, so that rate can range from $0 \to \infty$

$$\begin{aligned}
rate &= r \text{ (events per unit exposure)} \\
Count &= C \text{ (Number of events)} \\
ExposureTime &= T \\
C &\sim \text{poisson}(rT) \\
E[C] &= rT
\end{aligned}$$

The Poisson Regression Model

$$
\begin{aligned}
\log(\hat{r}) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \\
\hat{r} &= e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} \\
E[C] &= Tr \\
&= T \times e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} \\
&= e^{\log(T) + \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p} \\
log(E[C]) &= \log(T) + \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p
\end{aligned}
$$

Introduction
Poisson Regression
Negative Binomial Regression
Additional topics
Introduction
Example
Goodness of Fit
Constraints
Other considerations

Parameter Interpretation

- When x_i increases by 1, $\log(r)$ increases by β_i
- Therefore, r is multiplied by e^{β_i}
- As with logistic regression, coefficients are less interesting than their exponents
- e^{β} is the Incidence Rate Ratio

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Introduction
Poisson Regression
Negative Binomial Regression
Additional topics
Introduction
Example
Goodness of Fit
Constraints
Other considerations

Poisson Regression in Stata

- Command `poisson` will do Poisson regression
- Enter the exposure with the option `exposure(`*`varname`*`)`
- Can also use `offset(`*`lvarname`*`)`, where *`lvarname`* is the log of the exposure
- To obtain Incidence Rate Ratios, use the option `irr`

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Introduction
Poisson Regression
Negative Binomial Regression
Additional topics

Poisson Regression Example: Doctor's Study

	Smokers		Non-smokers	
Age	Deaths	Person-Years	Deaths	Person-Years
35–44	32	52,407	2	18,790
45–54	104	43,248	12	10,673
55–64	206	28,612	28	5,710
65–74	186	12,663	28	2,585
75–84	102	5,317	31	1,462

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

```
. poisson deaths i.agecat i.smokes, exp(pyears) irr

Poisson regression                              Number of obs   =         10
                                                LR chi2(5)      =     922.93
                                                Prob > chi2     =     0.0000
Log likelihood = -33.600153                     Pseudo R2       =     0.9321

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
      45-54  |   4.410584   .8605197     7.61   0.000     3.009011    6.464997
      55-64  |    13.8392   2.542638    14.30   0.000     9.654328    19.83809
      65-74  |   28.51678   5.269878    18.13   0.000     19.85177    40.96395
      75-84  |   40.45121   7.775511    19.25   0.000     27.75326    58.95885
             |
      smokes |
        Yes  |   1.425519   .1530638     3.30   0.001     1.154984    1.759421
       _cons |   .0003636   .0000697   -41.30   0.000     .0002497    .0005296
  ln(pyears) |          1  (exposure)
------------------------------------------------------------------------------

.
```


Using `predict` after `poisson`

Options available:

`n`	(default)	expected number of events (rate × duration of exposure)
`ir`		incidence rate
`xb`		linear predictor

Example: `predict`

`predict pred_n`

	Smokers		Non-smokers	
Age	Deaths	`pred_n`	Deaths	`pred_n`
35–44	32	27.2	2	6.8
45–54	104	98.9	12	17.1
55–64	206	205.3	28	28.7
65–74	186	187.2	28	26.8
75–84	102	111.5	31	21.5

Goodness of Fit

- Command `estat gof` compares observed and expected (from model) counts
- Can detect whether the Poisson model is reasonable
- If not could be due to
  - Systematic part of model poorly specified
  - Random variation not really Poisson
- Degrees of freedom for test = number of categories of observations - number of coefficients in model (including `_cons`)

Goodness of Fit Example

```
. estat gof

         Deviance goodness-of-fit =  12.13244
         Prob > chi2(4)           =    0.0164

         Pearson goodness-of-fit  =  11.15533
         Prob > chi2(4)           =    0.0249
```


Improving the fit of the model

- If the model fit is poor, it can be improved by:
  - Allowing for non-linearity of associations
  - Introducing interaction terms
  - Including other variables

Example: Improving fit of the model

```
. poisson deaths i.agecat##i.smokes, exp(pyears) irr

Poisson regression                                Number of obs   =         10
                                                  LR chi2(9)      =     935.07
                                                  Prob > chi2     =     0.0000
Log likelihood =  -27.53397                       Pseudo R2       =     0.9444

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
      45-54  |    10.5631   8.067701     3.09   0.002     2.364153    47.19623
      55-64  |   46.07004   33.71981     5.23   0.000     10.97496    193.3901
      65-74  |    101.764   74.48361     6.32   0.000     24.24256    427.1789
      75-84  |   199.2099   145.3356     7.26   0.000     47.67693    832.3648
             |
      smokes |
        Yes  |   5.736637   4.181256     2.40   0.017     1.374811    23.93711
             |
agecat#smokes |
   45-54#Yes  |   .3728337   .2945619    -1.25   0.212     .0792525    1.753951
   55-64#Yes  |   .2559409   .1935392    -1.80   0.072     .0581396    1.126697
   65-74#Yes  |   .2363859   .1788334    -1.91   0.057     .0536612    1.041316
   75-84#Yes  |   .1577109   .1194146    -2.44   0.015     .0357565    .6956154
             |
       _cons |   .0001064   .0000753   -12.94   0.000     .0000266    .0004256
   ln(pyears) |          1  (exposure)
------------------------------------------------------------------------------
```



```
. testparm i.agecat#i.smokes

           chi2(  4) =   10.20
         Prob > chi2 =    0.0372

. lincom 1.smokes + 5.age#1.smokes, eform

 ( 1)  [deaths]1.smokes + [deaths]5.agecat#1.smokes = 0

------------------------------------------------------------------------------
      deaths |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9047304   .1855513    -0.49   0.625     .6052658     1.35236
------------------------------------------------------------------------------

. estat gof

         Deviance goodness-of-fit =  .0000694
         Prob > chi2(0)           =         .

         Pearson goodness-of-fit  =  1.14e-13
         Prob > chi2(0)           =         .
```

Constraints

- Can force parameters to be equal to each other or specified value
- Can be useful in reducing the number of parameters in a model
- Simplifies description of model
- Enables goodness of fit test
- Syntax: `constraint define` *`n varname`* `=` *`expression`*

Constraint Example

```
. constraint define 1 3.agecat#1.smokes = 4.agecat#1.smokes

. poisson deaths i.agecat##i.smokes, exp(pyears) irr constr(1)

Poisson regression                                Number of obs   =         10
                                                  Wald chi2(8)    =     632.14
Log likelihood = -27.572645                       Prob > chi2     =     0.0000

 ( 1)  [deaths]3.agecat#1.smokes - [deaths]4.agecat#1.smokes = 0
------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
      45-54  |    10.5631   8.067701     3.09   0.002     2.364153    47.19623
      55-64  |     47.671   34.37409     5.36   0.000     11.60056    195.8978
      65-74  |   98.22765   70.85012     6.36   0.000     23.89324    403.8244
      75-84  |   199.2099   145.3356     7.26   0.000     47.67693    832.3648
             |
      smokes |
        Yes  |   5.736637   4.181256     2.40   0.017     1.374811    23.93711
             |
agecat#smokes |
   45-54#Yes  |   .3728337   .2945619    -1.25   0.212     .0792525    1.753951
   55-64#Yes  |   .2461772    .182845    -1.89   0.059     .0574155    1.055521
   65-74#Yes  |   .2461772    .182845    -1.89   0.059     .0574155    1.055521
   75-84#Yes  |   .1577109   .1194146    -2.44   0.015     .0357565    .6956154
             |
       _cons |   .0001064   .0000753   -12.94   0.000     .0000266    .0004256
  ln(pyears) |          1  (exposure)
------------------------------------------------------------------------------
```


Constraint Example Cont.

```
. estat gof

         Deviance goodness-of-fit =  .0774185
         Prob > chi2(1)           =    0.7808

         Pearson goodness-of-fit  =  .0773882
         Prob > chi2(1)           =    0.7809
```


Predicted Numbers from Poisson Regression Model

	Smokers			Non-smokers		
Age	Observed	Pred 1	Pred 2	Observed	Pred 1	Pred 2
35–44	32	27.2	32.0	2	6.8	2.0
45–54	104	98.9	104.0	12	17.1	12.0
55–64	206	205.3	205.0	28	28.7	29.0
65–74	186	187.2	187.0	28	26.8	27.0
75–84	102	111.5	102.0	31	21.5	31.0

Pred 1 No Interaction

Pred 2 Interaction & Constraint

Zeros

- May be structural (Exposure = 0, so count *had* to be 0)
- Don't count towards DOF
- Lead to problems in estimation
  - IRR is huge or tiny
  - SE is huge
  - Confidence interval is undefined
- Stata may be unable to produce a confidence interval

Overdispersion

- Adding predictors to model may not lead to an adequate fit
- There may be variation between individuals in rate not included in model
- Variance is equal to mean for a Poisson distribution
- The variation between individuals means there is more variation than expected: overdispersion
- If there is overdispersion, standard errors will be too small

Negative Binomial Regression

- Allows for extra variation
- Assumes a mixture of Poisson variables, with the means having a given distribution
- Two possible models:
  - Var(Y) = $\mu(1 + \delta)$
  - Var(Y) = $\mu(1 + \alpha\mu)$
- α or δ is the overdispersion parameter
- $\alpha = 0$ or $\delta = 0$ gives the Poisson model.

Negative Binomial Regression in Stata

- Command `nbreg`
- Syntax similar to `poisson`
- Default gives Var(Y) = $\mu(1 + \alpha\mu)$
- Option `dispersion(constant)` gives Var(Y) = $\mu(1 + \delta)$

Negative Binomial Regression Example

```
. poisson deaths i.cohort, exposure(exposure) irr

Poisson regression                              Number of obs   =         21
                                                LR chi2(2)      =      49.16
                                                Prob > chi2     =     0.0000
Log likelihood = -2159.5158                     Pseudo R2       =     0.0113

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cohort |
  1960-1967  |   .7393079   .0423859    -5.27   0.000     .6607305      .82723
  1968-1976  |   1.077037   .0635156     1.26   0.208      .959474    1.209005
             |
       _cons |   .0202523   .0008331   -94.80   0.000     .0186836    .0219527
ln(exposure) |          1  (exposure)
------------------------------------------------------------------------------

. estat gof

         Deviance goodness-of-fit =  4190.689
         Prob > chi2(18)          =    0.0000

         Pearson goodness-of-fit  =  15387.67
         Prob > chi2(18)          =    0.0000
```



```
. nbreg deaths i.cohort, exposure(exposure) irr

Negative binomial regression                    Number of obs   =         21
                                                LR chi2(2)      =       0.40
Dispersion     = mean                           Prob > chi2     =     0.8171
Log likelihood = -131.3799                      Pseudo R2       =     0.0015

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cohort |
  1960-1967  |   .7651995   .5537904    -0.37   0.712     .1852434    3.160869
  1968-1976  |   .6329298   .4580292    -0.63   0.527     .1532395    2.614209
             |
       _cons |   .1240922   .0635173    -4.08   0.000     .0455042    .3384052
ln(exposure) |          1  (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |   .5939963   .2583615                       .087617    1.100376
-------------+----------------------------------------------------------------
       alpha |   1.811212   .4679475                       1.09157    3.005294
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 4056.27 Prob>=chibar2 = 0.000
```


Log-Linear Models

- An $R \times C$ table is simply a series of counts
- The counts have two predictor variables (rows and columns)
- Can fit a Poisson model to such a table
- Association between two variables is given by the interaction between the variables
- Model: $\log(p) = \beta_0 + \beta_r x_r + \beta_c x_c + \beta_{rc} x_{rc}$
- For a 2×2 table, such a model is *exactly* equivalent to logistic regression.

Log-Linear Modelling Example

Outcome	Exposure	
	Exposed	Unexposed
Cases	20	10
Non-cases	10	20

OR = 4

Log-linear modelling example: stata output

```
     +---------------------------+
     | outcome   exposure   freq |
     |---------------------------|
  1. |       0          0     20 |
  2. |       1          0     10 |
  3. |       0          1     10 |
  4. |       1          1     20 |
     +---------------------------+

. xi: poisson freq i.exp*i.out, irr

Poisson regression                                Number of obs   =          4
                                                  LR chi2(3)      =       6.80
                                                  Prob > chi2     =     0.0787
Log likelihood = -8.9990653                       Pseudo R2       =     0.2741

------------------------------------------------------------------------------
        freq |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iexposure_1 |         .5   .1936492    -1.79   0.074     .2340459    1.068166
 _Ioutcome_1 |         .5   .1936492    -1.79   0.074     .2340459    1.068166
_IexpXout_~1 |          4    2.19089     2.53   0.011     1.367218    11.7026
------------------------------------------------------------------------------

. logistic outcome exposure [fw=freq]

Logistic regression                               Number of obs   =         60
                                                  LR chi2(1)      =       6.80
                                                  Prob > chi2     =     0.0091
Log likelihood =  -38.19085                       Pseudo R2       =     0.0817

------------------------------------------------------------------------------
     outcome | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |          4    2.19089     2.53   0.011     1.367218    11.7026
------------------------------------------------------------------------------
```


Direct & Indirect Standardisation

- Used for comparing rates between populations
- Assumes covariates differ between populations
- What would rates be if the covariates were the same ?
  - I.e. same proportion of subjects in each stratum
  - Proportions from standard population = direct standardisation
  - Proportions from this population = indirect standardisation

Direct Standardisation

- Calculate rate in each stratum
- Standardised rate = weighted mean of these rates
- Weights = proportions of subjects in each stratum of standard population.
- Standardised rate = what rate would be in standard population if it had the same stratum specific rates as our population
- Different standard = different standardised rate
- Can compare directly adjusted rates (adjusted to same population)

Indirect Standardisation

- Per stratum rates are unavailable/unreliable
- Use known rates from a standard population
- Weight known rates according to stratum size our population
- Produce expected number of events if standard rates apply
- Ratio $\frac{Observed}{Expected}$ = SMR

Standardisation vs. Adjustment

- Direct standardisation
  - Poisson regression assumes same RR in each stratum
  - D.S. assumes different RR in each stratum
  - Both give weighted mean RR: weights differ
- Indirect Standardisation
  - Good measure of causal effect in this sample
  - Can be useful in e.g. observational study of treatment effect.
  - Do not compare SMR's
    - They tell you what happened in observed group.
    - Do not tell you what might happen in a different group.

Generalized Linear Models

- We have met a number of regression models
- All have the form:

$$
\begin{aligned}
g(\mu) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \\
Y &= \mu + \varepsilon
\end{aligned}
$$

where μ is the expected value of Y
ε has a known distribution (normal, binomial etc)
$g()$ is called the link function

Components of a GLM

- You can choose the link function for yourself
- It should:
  - Map $-\infty$ to ∞ onto reasonable values for μ
  - Have parameters that are easy to interpret
- Error distribution is determined by the data
- Only certain distributions are allowed

Examples of GLM's

Model	Range of μ	Link		Error Distribution
Linear Regression	$-\infty$ to ∞	$g(\mu)$	$= \mu$	Normal
Logistic Regression	0 to 1	$g(\mu)$	$=\log(\frac{\mu}{1-\mu})$	Binomial
Poisson Regression	0 to ∞	$g(\mu)$	$=\log(\mu)$	Poisson

GLM's in Stata

- Command `glm`
- Option `family()` sets the error distribution
- Option `link()` sets the link function
- There are more options to `predict` after `glm`

E.g.	`glm` *`yvar xvars`*`, family(binomial) link(logit)`
is equivalent to	`logistic` *`yvar xvars`*

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Introduction
Poisson Regression
Negative Binomial Regression
Additional topics
Log-linear Models
Standardisation
Generalized Linear Models
Setting Reference Category for Categorical Variables

Setting Reference Category for Categorical Variables: Old Way

`char` *`variable`*`[omit] #`

`char`	Characteristic
`variable`	Name of variable to set reference category for
`#`	Value of reference category

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Introduction
Poisson Regression
Negative Binomial Regression
Additional topics
Log-linear Models
Standardisation
Generalized Linear Models
Setting Reference Category for Categorical Variables

Setting Reference Category for Categorical Variables: New Way

For one model	`ib#.`*`varname`*
Permanently	`fvset base #` *`varname`*
Alternatives to `#`	`first`
	`last`
	`frequent`

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS