***Stata tips #5 (STATA session#5)***

***Epidemiologic Methods***

***Adapted from notes of previous TAs***

This week:

1. Using and interpreting Poisson regression using STATA

# Poisson regression:

* + Counted outcomes
    - Number of sick persons
    - Number of moles on body
  + Can also be a rate if specify time observed, i.e. exposure time
    - Number of c-sections performed at a hospital in a certain year
    - The number of Olympic medals won by a country this year
  + Can also be a proportion if specify population observed (here we use mid-point population x 1 year, so it is a rate essentially)
    - Cancer cases per age group in a population
  + Some assumption of Poission regression:
    - Log of the outcome varies linearly with each unit increase in the predictor variable, i.e. it is loglinear
    - The variance equals the mean
    - Observations are independent

The Stata command and interpretation:

poisson:

* + - predictor coefficients – the change in log count (or log rate, or log proportion) for each unit change in the predictor. (Can add the irr option to get incidence rate ratio if outcome is incidence rate)
    - \_cons – the log count (or log incidence rate) given that all predictor variables are zero (at baseline). When adding the irr option, get the average count or incidence rate for baseline predictors.
    - Use the option exposure()in the command when want to specify length of time you can observe the outcome (or number of people in which you observe the outcome). Represented in Stata output as ln(exposure), and is called an offset. This is always constrained to 1. For an intuitive explanation of this: [http://stats.stackexchange.com/questions/12005/interpretation-of-](http://stats.stackexchange.com/questions/12005/interpretation-of-intercept-term-in-poisson-model-with-offset-and-covariates)

[intercept-term-in-poisson-model-with-offset-and-covariates](http://stats.stackexchange.com/questions/12005/interpretation-of-intercept-term-in-poisson-model-with-offset-and-covariates)

In fact, the option exposure()is added in almost all Poisson regressions.

* + - You can also use the option offset(), and put log(person-year) in the

().

We want to answer the following question: *What is the association between serum glucose levels and stroke outcome?*

After performing some basic variable generation and data cleaning, we can run some basic Poisson regressions calculating the different types of incidences.

First let’s review what the definition of incidence is, and what are the measures of incidence?

* + **cumulative incidence: events/total pop** within a time period (a proportion)
  + **incidence rate: events/given time period** (a rate used if individuals all have equal follow-up)
  + **Incidence density (also referred to as incidence rate): events/person-time** (a rate used if have specific individual follow-up time)

Using poisson regression to calculate **cumulative incidence**:

| . poisson stroke1, nolog | |  | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Poisson regression | | Number of obs | | | | = | 4402 |
|  | | LR chi2(0) | | | | = | -0.00 |
|  | | Prob > chi2 | | | | = | . |
| Log likelihood = -1318.2015 | | Pseudo R2 | | | | = | -0.0000 |
| stroke1 | | Coef. | Std. Err. | z | P>|z| | [95% Conf. | | Interval] |
| + |  |  |  |  |  | |  |
| \_cons | | -2.441779 | .0510976 | -47.79 | 0.000 | -2.541929 | | -2.34163 |

**Here we assume that each individual has equal units of exposure**. Is this true? The \_cons represents the log number of strokes per person.

| . poisson, irr | |  | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Poisson regression | | Number of obs | | | | = | 4402 |
|  | | LR chi2(0) | | | | = | -0.00 |
|  | | Prob > chi2 | | | | = | . |
| Log likelihood = -1318.2015 | | Pseudo R2 | | | | = | -0.0000 |
| stroke1 | | IRR | Std. Err. | z | P>|z| | [95% Conf. | | Interval] |
| + |  |  |  |  |  | |  |
| \_cons | | .0870059 | .0044458 | -47.79 | 0.000 | .0787144 | | .0961708 |

On the rate ratio scale, the \_cons represents the number of strokes per person. Stata assumes default exposure(1).

Let’s express exposure variable in 1000 person-years

generate strkexp=timestrk1/(365.25\*1000)

Next, we want to estimate **incidence rate (density)** with our strkexp variable:

| . poisson stroke1, exposure(strkexp) | nolog |  | |
| --- | --- | --- | --- |
|  |  |  | |
| Poisson regression |  | Number of obs = | 4402 |
|  |  | LR chi2(0) = | 0.00 |
|  |  | Prob > chi2 = | . |
| Log likelihood = -1531.5148 |  | Pseudo R2 = | 0.0000 |
| stroke1 | Coef. Std. Err. | z | P>|z| [95% Conf. | Interval] |
| + |  |  |  |
| \_cons | 1.464232 .0510976 | 28.66 | 0.000 1.364082 | 1.564381 |
| ln(strkexp) | 1 (exposure) |  |  |  |

Here we are modelling ‘stroke per thousand person years’ and the \_cons represents the log number of strokes per 1000-person years (why per 1000-person years? Because the unit of strkexp is set to 1000 years).

The ln(strkexp) coefficient is just 1 and this is by definition. Its presence changes our interpretation of the intercept and any additional predictor variable coefficients in the model.

# Now let’s investigate the effect of blood glucose level on strokes:

To increase interpretability of beta0, let us center glucose1 at 100mg/dL (median normal range)

gen centglucose=glucose1-100

. poisson stroke1 centglucose, exposure(strkexp) nolog

| Poisson regression | | | Number of obs | | | | = | 4007 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | LR chi2(1) | | | | = | 14.15 |
|  | | | Prob > chi2 | | | | = | 0.0002 |
| Log likelihood = -1399.0464 | | | Pseudo R2 | | | | = | 0.0050 |
| stroke1  centglucose | |  +  | | Coef.  .0068862 | Std. Err.  .0015159 | z  4.54 | P>|z|  0.000 | [95% Conf.  .0039151 | | Interval]  .0098574 |
| \_cons | | | 1.596445 | .0573668 | 27.83 | 0.000 | 1.484008 | | 1.708882 |
| ln(strkexp) | | | 1 | (exposure) |  |  |  | |  |

. poisson, irr

Poisson regression Number of obs = 4007

|  | | | LR chi2(1) | | | | = | 14.15 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Prob > chi2 | | | | = | 0.0002 |
| Log likelihood | = | -1399.0464 | Pseudo R2 | | | | = | 0.0050 |
| stroke1 |  +  centglucose | | IRR  1.00691 | | Std. Err.  .0015264 | z  4.54 | P>|z|  0.000 | [95% Conf.  1.003923 | | Interval]  1.009906 |
| \_cons | | 4.935454 | | .2831313 | 27.83 | 0.000 | 4.410587 | | 5.522781 |
| ln(strkexp) | | 1 | | (exposure) |  |  |  | |  |

What’s the interpretation of the coefficient for centglucose?

There is a 0.7% increase in incidence rate of stroke for each unit increase (mg/dL) in blood glucose level

What’s the interpretation of \_cons?

| Is age a confounder? | | | |  | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| . poisson stroke1 centglucose | | | | i.agecat, | exposure(strkexp) nolog | | |  |  |
| Poisson regression | | | |  | Number of obs | | | = | 4007 |
|  | | | |  | LR chi2(4) | | | = | 203.65 |
|  | | | |  | Prob > chi2 | | | = | 0.0000 |
| Log likelihood = -1304.2973 | | | |  | Pseudo R2 | | | = | 0.0724 |
| stroke1 | | | Coef. | Std. Err. | | z | P>|z| | [95% | Conf. | Interval] |
| centglucose  agecat | |  |  | | .0043033 | .0016303 | | 2.64 | 0.008 | .001108 | | .0074987 |
| 45 | | | .8797112 | .1882664 | | 4.67 | 0.000 | .5107158 | | 1.248707 |
| 55 | | | 1.863246 | .1749809 | | 10.65 | 0.000 | 1.52029 | | 2.206202 |
| 65  \_cons | |  |  | | 2.255635  .4070698 | .2356501  .1615045 | | 9.57  2.52 | 0.000  0.012 | 1.79377  .0905269 | | 2.717501  .7236127 |
| ln(strkexp) | | | 1 | (exposure) | | | | | | |

4.93 is the incidence rate of stroke for someone with the blood glucose level 100mg/dl (blood glucose is centered to 100 mg/dl)

. poisson stroke1 centglucose i.agecat, exposure(strkexp) irr nolog

| Poisson regression | | | Number of obs | | | | = | 4007 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | LR chi2(4) | | | | = | 203.65 |
|  | | | Prob > chi2 | | | | = | 0.0000 |
| Log likelihood = -1304.2973 | | | Pseudo R2 | | | | = | 0.0724 |
| stroke1  centglucose | |  +  | | IRR  1.004313 | Std. Err.  .0016373 | z  2.64 | P>|z|  0.008 | [95% Conf.  1.001109 | | Interval]  1.007527 |
| agecat | |  | |  |  |  |  |  | |  |
| 45 | | | 2.410204 | .4537604 | 4.67 | 0.000 | 1.666484 | | 3.485832 |
| 55 | | | 6.444622 | 1.127686 | 10.65 | 0.000 | 4.57355 | | 9.081162 |
| 65  \_cons | |  |  | | 9.541353  1.502409 | 2.248421  .2426457 | 9.57  2.52 | 0.000  0.012 | 6.012073  1.094751 | | 15.14243  2.061869 |
| ln(strkexp) | | | 1 | (exposure) | | | | | |

How about smoking?

. poisson stroke1 centglucose i.agecat cursmoke1, exposure(strkexp) nolog Poisson regression Number of obs = 4007

LR chi2(5) = 212.16

|  | | | | Prob > | | | chi2 | = | 0.0000 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Log likelihood | | = | -1300.0401 | Pseudo | | | R2 | = | 0.0754 |
| stroke1 | | | Coef. | | Std. Err. | z | P>|z| | [95% Conf. | | Interval] |
|  | + |  | |  |  |  |  | |  |
| centglucose | | | .004532 | | .0016355 | 2.77 | 0.006 | .0013265 | | .0077375 |
|  | | |  | |  |  |  |  | |  |
| agecat | | |  | |  |  |  |  | |  |
| 45 | | | .9107819 | | .1885469 | 4.83 | 0.000 | .5412367 | | 1.280327 |
| 55 | | | 1.938613 | | .1768412 | 10.96 | 0.000 | 1.592011 | | 2.285216 |
| 65 | | | 2.353381 | | .2380121 | 9.89 | 0.000 | 1.886886 | | 2.819876 |
|  | | |  | |  |  |  |  | |  |
| cursmoke1 | | | .3181767 | | .1086589 | 2.93 | 0.003 | .1052092 | | .5311441 |
| \_cons | | | .2113547 | | .1763226 | 1.20 | 0.231 | -.1342314 | | .5569407 |
| ln(strkexp) | | | 1 | | (exposure) |  |  |  | |  |

. poisson, irr

Poisson regression Number of obs = 4007

|  | | | | LR chi2(5) | | | | = | 212.16 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Prob > chi2 | | | | = | 0.0000 |
| Log likelihood | | = | -1300.0401 | Pseudo R2 | | | | = | 0.0754 |
| stroke1 | | | IRR | | Std. Err. | z | P>|z| | [95% Conf. | | Interval] |
|  | + |  | |  |  |  |  | |  |
| centglucose | | | 1.004542 | | .0016429 | 2.77 | 0.006 | 1.001327 | | 1.007768 |
|  | | |  | |  |  |  |  | |  |
| agecat | | |  | |  |  |  |  | |  |
| 45 | | | 2.486266 | | .4687778 | 4.83 | 0.000 | 1.71813 | | 3.597816 |
| 55 | | | 6.949109 | | 1.228889 | 10.96 | 0.000 | 4.91362 | | 9.827806 |
| 65 | | | 10.52108 | | 2.504144 | 9.89 | 0.000 | 6.598786 | | 16.77477 |
|  | | |  | |  |  |  |  | |  |
| cursmoke1 | | | 1.374619 | | .1493646 | 2.93 | 0.003 | 1.110943 | | 1.700877 |
| \_cons | | | 1.23535 | | .2178202 | 1.20 | 0.231 | .8743877 | | 1.745325 |
| ln(strkexp) | | | 1 | | (exposure) |  |  |  | |  |

# How about gender?

. poisson stroke1 centglucose i.agecat cursmoke1 female, exposure(strkexp) nolog

Poisson regression Number of obs = 4007 LR chi2(6) = 215.27

|  | |  |  | Prob > | | | chi2 | = | 0.0000 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Log likelihood | | = | -1298.485 | Pseudo | | | R2 | = | 0.0765 |
| stroke1 |  +  centglucose | | | Coef.  .0044905 | | Std. Err.  .0016288 | z  2.76 | P>|z|  0.006 | [95% Conf.  .001298 | | Interval]  .0076829 |
| | agecat |  45 | | | .9132776 | | .1885318 | 4.84 | 0.000 | .543762 | | 1.282793 |
| 55 | | | 1.937433 | | .1767258 | 10.96 | 0.000 | 1.591056 | 2.283809 | |
| 65  cursmoke1 | |  |  | | 2.347119  .2717249 | | .237882  .1117346 | 9.87  2.43 | 0.000  0.015 | 1.880879  .0527291 | 2.813359  .4907207 | |
| female | | | -.193771 | | .1096833 | -1.77 | 0.077 | -.4087463 | .0212044 | |
| \_cons | | | .3386517 | | .1892375 | 1.79 | 0.074 | -.032247 | .7095504 | |
| ln(strkexp) | | | 1 | | (exposure) |  |  |  |  | |

. poisson, irr

Poisson regression Number of obs = 4007

|  | | | | LR chi2(6) | | | | = | 215.27 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Prob > chi2 | | | | = | 0.0000 |
| Log likelihood | | = | -1298.485 | Pseudo R2 | | | | = | 0.0765 |
| stroke1 |  +  centglucose | | | IRR  1.004501 | | Std. Err.  .0016362 | z  2.76 | P>|z|  0.006 | [95% Conf.  1.001299 | | Interval]  1.007713 |
| | agecat |  45 | | | 2.492478 | | .4699114 | 4.84 | 0.000 | 1.722475 | | 3.6067 |
| 55 | | | 6.940908 | | 1.226638 | 10.96 | 0.000 | 4.908932 | 9.81399 | |
| 65  cursmoke1 | |  |  | | 10.4554  1.312226 | | 2.487152  .146621 | 9.87  2.43 | 0.000  0.015 | 6.559265  1.054144 | 16.6658  1.633493 | |
| female | | | .8238466 | | .0903622 | -1.77 | 0.077 | .6644828 | 1.021431 | |
| \_cons | | | 1.403055 | | .2655105 | 1.79 | 0.074 | .9682674 | 2.033077 | |
| ln(strkexp) | | | 1 | | (exposure) |  |  |  |  | |

It seems that glucose-stroke association is confounded mainly by age but not by other factors.