# Poisson Regression for Count Data

Bowei Kang

Department of Public Health Sciences
The University of Chicago

$\boxed{\text{NOT IN FINAL}}$

Adapted from Dr. Lin Chen's slides for PBHS 32700, Spring 2023.

# Poisson regression in GLM framework

| Response | Link function | Error | Model |
|---|---|---|---|
| Continuous | Identity | Normal | Linear |
| Binary | Logit | Binomial | Logistic |
| Categorical | Logit | Multinomial | Multinomial logistic |
| Counts | Natural log | Poisson | Poisson |

# Count data

- Count: the number of occurrences of an event (success) in a fixed interval of time and/or a specified region of space.
- The total number of such events is a non-negative integer but could be large.

  For example:
- The number of new cases of tuberculosis in ZIP code 60637 during a randomly-selected year.
- The number of accidents at a given intersection on a randomly-selected weekend.
- The number of epileptic seizures for a randomly-selected epilepsy patient over a 2 week period.

# Can we analyze count data with linear regression?

Treat counts as continuous, normally distributed?

- Count distribution is too skewed to satisfy normality (incorrect test results).
- Normal model does not necessarily prevent negative estimated counts.

Exception :

- If count is very large
- Don't care about interpretation of $\beta$

Can we do log transformation on counts?

- Not ideal. The log of zero count is negative infinity and we will lose those data.

# Can we analyze count data with logistic regression?

Dichotomize counts?

- Choose a cutoff $c$, and for each count $Y_i$, generate a new binary variable, $Z_i = 0$ if $Y_i \le c$; $Z_i = 1$ if $Y_i > c$. Fit a logistic regression to $Z_i$.
- If $c = 0$, then $Z$ is the indicator whether the event happen or not.
- Loss of $\overset{x}{\text{information}}$ resulting in under-powered tests. Is 1 event really equal to 100 events? Both $c=1$ and $c=100$ equal $Y=1$...

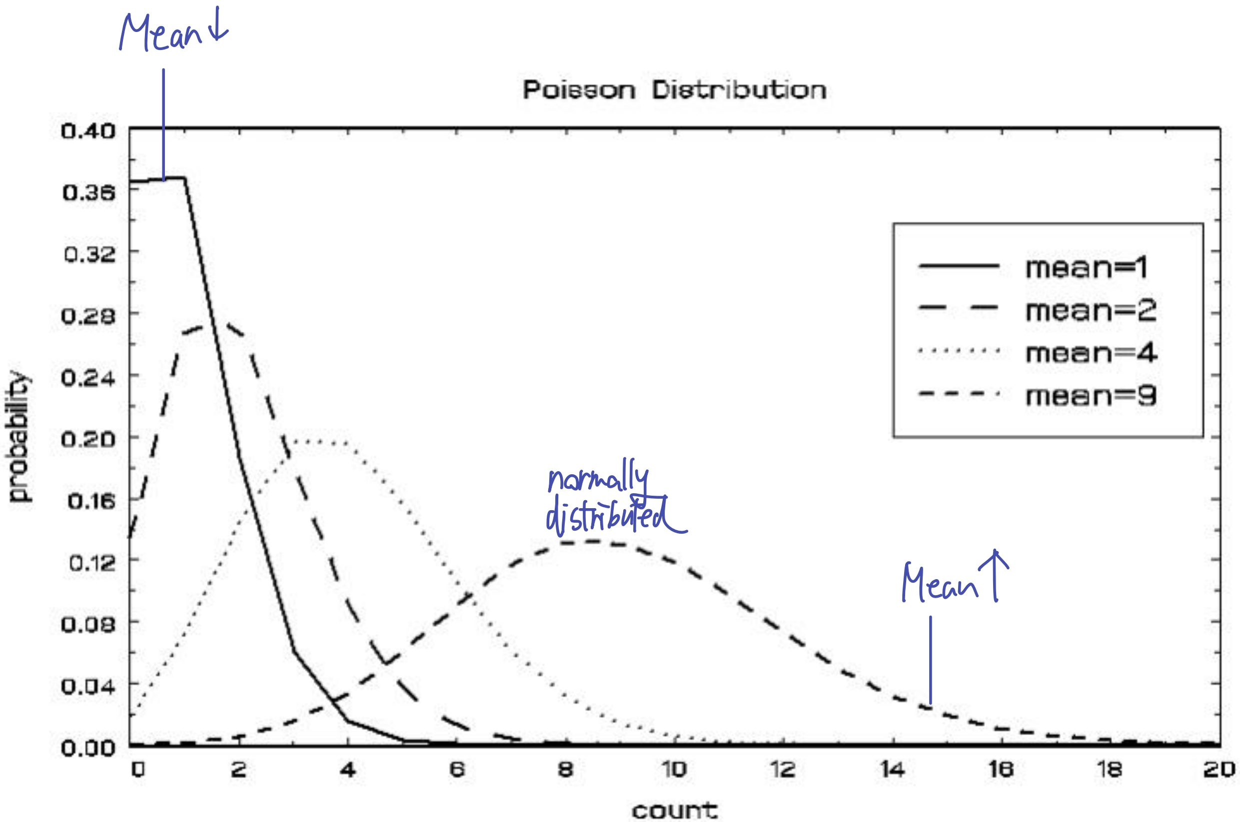
# Poisson Distribution

- A count variable is often assumed to follow Poisson distribution.
- Suppose that $Y$ is a count variable with probability function:

$$\Pr(Y = k) = \frac{e^{-\mu}\mu^k}{k!}, k = 0, 1, 2, ... \tag{1}$$

  Then $Y$ has a Poisson distribution with parameter $\mu$.
- $\mu$ is the mean number of events that occur in the given time interval and/or region of space.
- Basic assumptions: events occur independently with a known constant mean rate, $\mu$.
- A key property: $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$. Note that this is a property and could also be a restriction when assuming a variable following Poisson distribution.

# Poisson Distribution (cont.)




- The expected number of counts (per unit of time) is strictly positive.
- As mean increases, the probability at 0 decreases (shift to right); the distribution approximates normal.
- A larger mean correspond to a larger variance (more spread).

more spreading to the right

# Poisson Distribution: Examples

- $Y$ is the number of new cases of tuberculosis in ZIP code 60637 during a randomly-selected year; $\mu$ is the average number of tuberculosis in ZIP code 60637 per year.
- $Y$ is the number of accidents at a given intersection on a randomly-selected weekend; $\mu$ is the average number of accidents at a given intersection per weekend.
- $Y$ is the number of epileptic seizures for a randomly-selected epilepsy patient over a 2 week period; $\mu$ is the average number of seizures per person over a two-week period.

# Genesis of / heuristic for Poisson distribution

- $\mu$ (average number of tuberculosis per year) is the annual average population in ZIP code 60637 $\times$ the probability of each person having a new case of tuberculosis.
- $\mu$ (average number of accidents per weekend) is the average number of cars through the intersection per weekend $\times$ the probability of each car having an accident.
- $\mu$ (average number of seizure per person for a two-week period) is the number of, say, minutes in a 2-week period $\times$ the probability of having a seizure in any given minute.
- Thus, to an approximation:

$$\mu = np$$

where $n$ is very very *large* and $p$ is very very *small*.

# The relationship between the Binomial and Poisson distribution

- The Binomial distribution tends toward the Poisson distribution as $\underset{\text{large}}{n \longrightarrow \infty}$, $\underset{\text{small}}{p \longrightarrow 0}$ and $np$ stays constant.
- The Poisson random variable with mean $\mu$ is approximately binomial with large $n$ and small $p$ such that $\mu = np$.

# Exposure and rate

- Just as with binomial data, where $p$ is the parameter of interest, $\lambda$, a rate parameter is usually of interest with Poisson count data.
- Suppose that $Y$ is a count of events that arise at a (incidence) rate of $\lambda$ per unit-time of exposure for an exposure period of $A$, so that $\mu = \lambda A$.
- In Epidemiology, $A$ is called person-time: the number of time units (usually years) contributed to the exposure by each person under observation.
- Can be expressed in days or months, etc., but typically person-years.
- 10 people followed for one year contribute 10 person-years, as does 1 person followed for 10 years.
- Person-time is used when persons are observed in the study for varying amounts of time. useful for different follow-up

# Exposure and rate: Example

- In the *tuberculosis* example, $\lambda$ could be the rate of new tuberculosis cases per person-year; then $A$ is the average number of people in ZIP 60637 over a year × 1 year.
- In the *accident* example, $\lambda$ could be the rate of accidents through the intersection per thousand car-weekend-day; then $A$ is the average number of cars/1000 through the intersection per weekend day × 2 days.
- In the *epileptic seizure* example, $\lambda$ could be the rate of seizures per person-hour; then $A$ is $2$ weeks × $7$ days × $24$ hours × 1 person.

# Poisson Regression Model for Incidence Rate

- Consider count data $Y_i$ which are Poisson, as a function of incidence rate $\lambda_i$ and exposure time $A_i$. Suppose the $i^{th}$ population are with covariates $x_{i1}, \cdots, x_{ik}$. Then, we have $Y_i \sim$ *Poisson* $(\mu_i)$, where $\mu_i = \lambda_i A_i$.
- A Poisson (log-linear) regression model for incidence rate $\lambda_i$ is

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \tag{2}$$

  We care more about $\lambda_i$ rather than $\mu_i$.
- $\beta_0$ is the baseline log incidence rate (i.e. log of event rate in a period of time), and $\exp(\beta_0)$ is the baseline incidence rate when all $x_1 = \cdots = x_k = 0$.
- $\beta_1$ is the log incidence rate ratio (i.e., difference in log incidence rate) when $X_1$ increase by 1 unit, adjusting for other covariates. $\exp(\beta_1)$ is the incidence rate ratio (IRR).

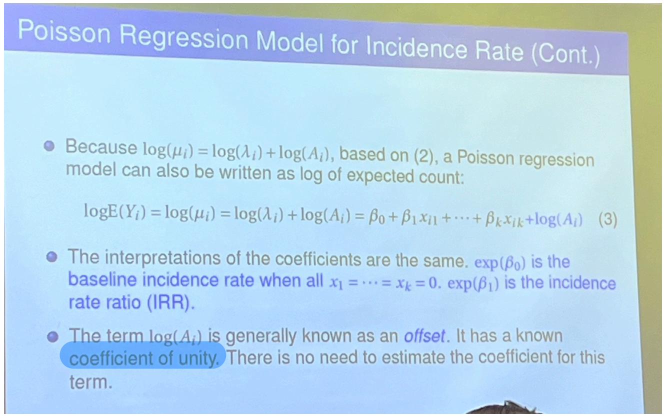
# Poisson Regression Model for Incidence Rate (Cont.)

- Because $\log(\mu_i) = \log(\lambda_i) + \log(A_i)$, based on (2), a Poisson regression model can also be written as log of expected count:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \underbrace{\log(A_i)}_{\text{constant}} \tag{3}$$

- The interpretations of the coefficients are the same. $\exp(\beta_0)$ is the baseline incidence rate when all $x_1 = \cdots = x_k = 0$. $\exp(\beta_1)$ is the incidence rate ratio (IRR). Interpret based on IRR "log IRR.."

- The mean/expected count $\mathrm{E}(Y_i) = \mu_i$ (not $\lambda_i$), thus we need a way to account for the exposure time $A_i$ in fitting (3).



# Fit a Poisson Regression Model

We do this with an *offset term* in the model equal to $\log(A_i)$.

- Offset is used to model rates per person-year, instead of just modeling the raw counts
- Offset is used to account for different group/population sizes, which could vary by age, region, other characteristics, etc.
- Offset *does not* have a $\beta$-coefficient associated with it, or, for which $\beta = 1$
- Without offset() options, exposure is assumed to be 1 for each subject (equivalent to assuming that exposure is unknown).

# Example: *British doctor's smoking and coronary death*

The data is from a very famous study where in 1951, all British doctors were sent a brief questionnaire about whether they smoked tobacco. Since then information about their deaths has been collected.

Table 1: Deaths from coronary heart disease after 10 years among British male doctors categorized by age and smoking status in 1951.

| Age | Smokers | | Non-smokers | |
|---|---|---|---|---|
| group | Deaths | Person-years | Deaths | Person-years |
| 35-44 | 32 | 52407 | 2 | 18790 |
| 45-54 | 104 | 43248 | 12 | 10673 |
| 55-64 | 206 | 28612 | 28 | 5710 |
| 65-74 | 186 | 12663 | 28 | 2585 |
| 75-84 | 102 | 5317 | 31 | 1462 |

Person-year is the sum of exposure years (years at risk or years in the study) for all subjects in the group. When a study subject develops the event (death) or leaves the study, they are no longer at risk and will no longer contribute person-year at risk.

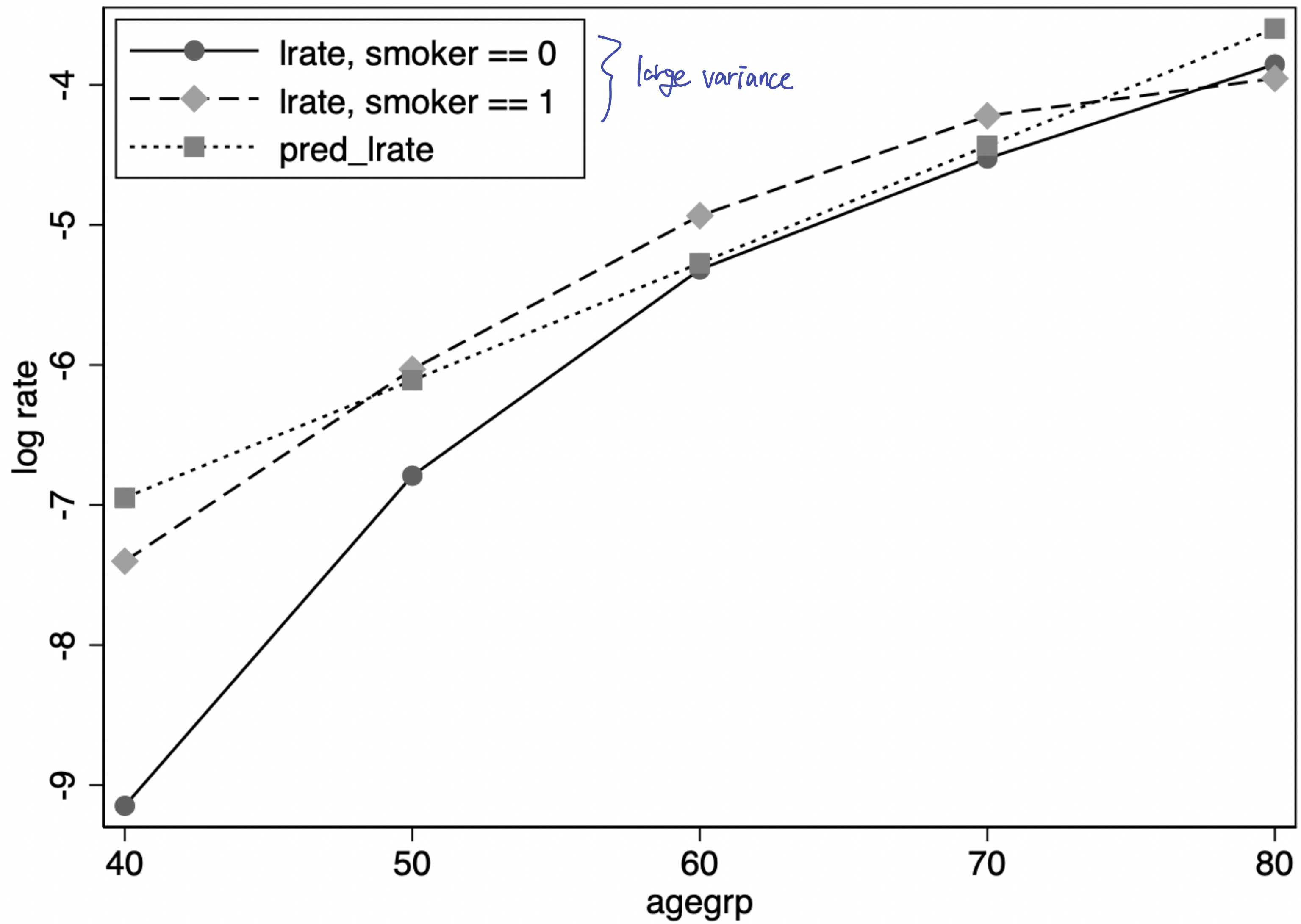
# British doctor's smoking and coronary death

```
. gen age = (agegrp-40)/10

* Take the midpoint of the age range, and generate a new variable age
  which denotes the number of decades from the 35-44 years group.

. list

     +------------------------------------------------+
     | agegrp   smoker   death   personyr   age |
     |------------------------------------------------|
  1. |     40        1      32      52407     0 |
  2. |     40        0       2      18790     0 |
  3. |     50        1     104      43248     1 |
  4. |     50        0      12      10673     1 |
  5. |     60        1     206      28612     2 |
     |------------------------------------------------|
  6. |     60        0      28       5710     2 |
  7. |     70        1     186      12663     3 |
  8. |     70        0      28       2585     3 |
  9. |     80        1     102       5317     4 |
 10. |     80        0      31       1462     4 |
     +------------------------------------------------+
```

# Model 1: Continuous age as the predictor

- We fit a Poisson regression model with `death` count as response and age as predictor. We set the variable `personyr` as "offset".
- $\log E(death_i) = \log(personyear_i) + \beta_0 + \beta_1 age_i$
- Note that the Stata code use "`exposure`" instead of "offset".
- $\beta_1$ represents incremental log death rate for every decade of age increase.

```
. poisson death age, exposure(personyr) nolog

Poisson regression                              Number of obs     =         10
                                                LR chi2(1)        =     850.06
                                                Prob > chi2       =     0.0000
Log likelihood =  -70.03973                     Pseudo R2         =     0.8585

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .8377632   .0288947    28.99   0.000     .7811305    .8943958
       _cons |   -6.94774   .0787198   -88.26   0.000    -7.102027   -6.793452
 ln(personyr)|          1  (exposure)
------------------------------------------------------------------------------
. poisgof

         Deviance goodness-of-fit =  85.01159
         Prob > chi2(8)           =    0.0000

         Pearson goodness-of-fit  =  75.24859
         Prob > chi2(8)           =    0.0000
```



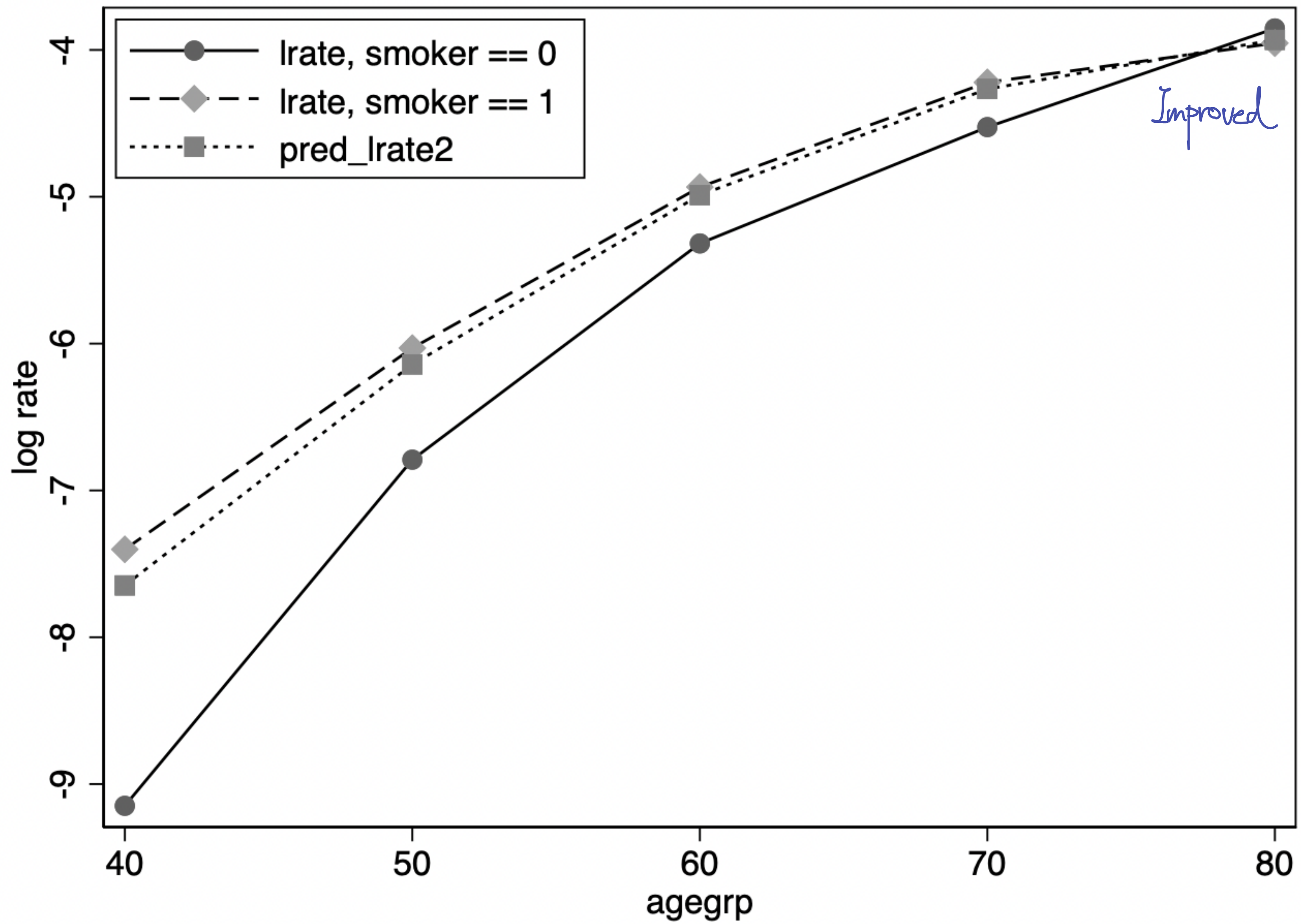
lrate, smoker == 0
lrate, smoker == 1
pred_lrate
large variance
log rate
-4
-5
-6
-7
-8
-9
40
50
60
70
80
agegrp

# Model 2: Categorical age as the predictor

- Model 1 has a large deviance.
- The log death rate increment is getting smaller as age increases.
- $\log E(death_i) =$ $\log(personyear_i) + \beta_0 + \beta_1 age_{1i} + \beta_2 age_{2i} + \beta_3 age_{3i} + \beta_4 age_{4i}$
- Use `i.age` to treat age as a categorical variable in regression.
- $\beta_4$ represents incremental log death rate for age group 75-84 versus baseline age group 35-44.

```
. poisson death i.age, exposure(personyr) nolog

Poisson regression                              Number of obs     =         10
                                                LR chi2(4)        =     911.08
                                                Prob > chi2       =     0.0000
Log likelihood = -39.528731                     Pseudo R2         =     0.9202

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
          1  |    1.50516   .1950191     7.72   0.000      1.12293    1.887391
          2  |   2.658625   .1835355    14.49   0.000     2.298902    3.018348
          3  |   3.380618   .1846203    18.31   0.000     3.018769    3.742467
          4  |    3.71561   .1921733    19.33   0.000     3.338957    4.092262
             |
       _cons |  -7.646845   .1714986   -44.59   0.000    -7.982976   -7.310714
ln(personyr) |          1  (exposure)
------------------------------------------------------------------------------
. poisgof

Deviance goodness-of-fit =  23.98959
Prob > chi2(5)           =    0.0002

Pearson goodness-of-fit  =     20.08
Prob > chi2(5)           =    0.0012
```

Better than Model 1.

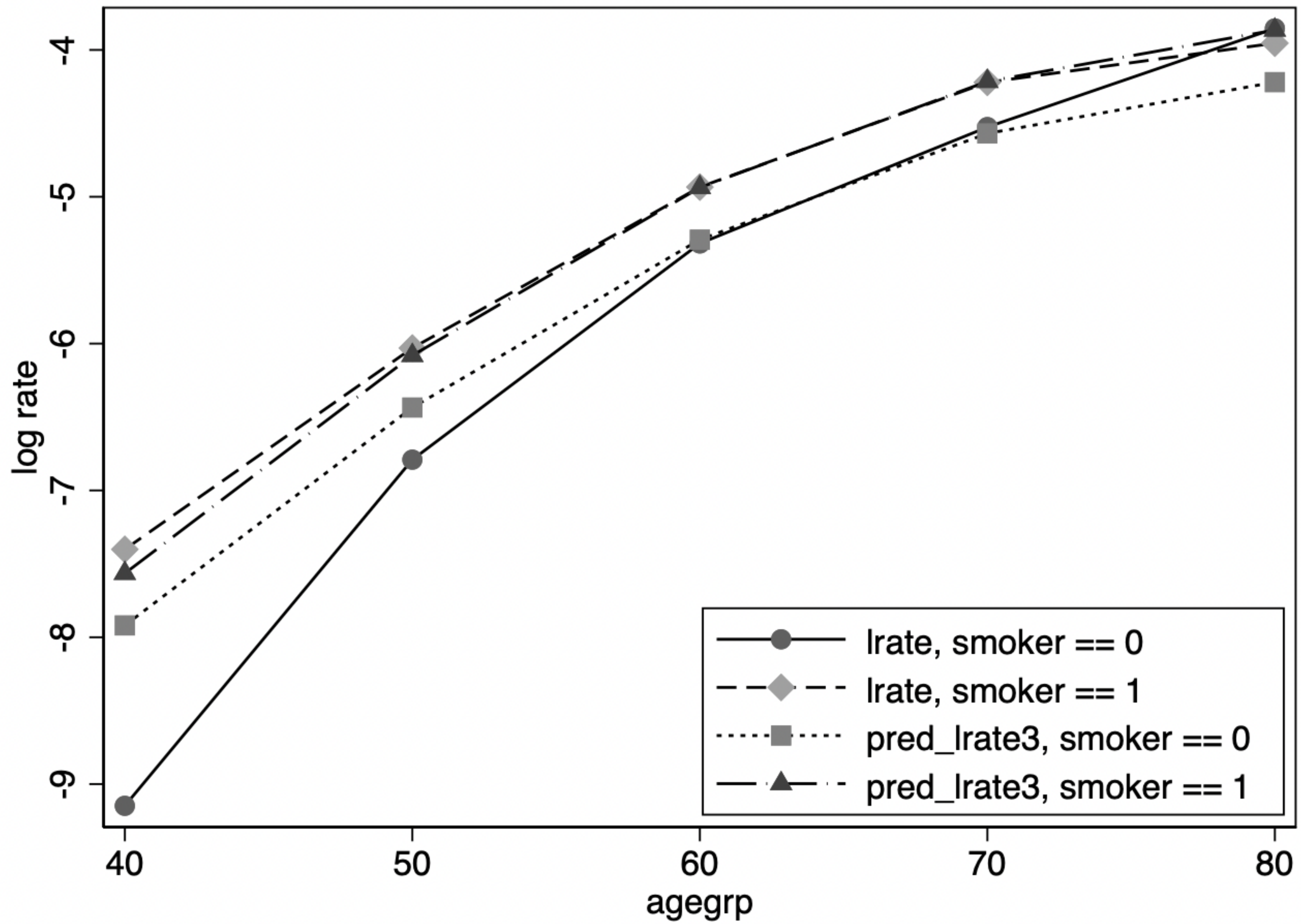

lrate, smoker == 0
lrate, smoker == 1
pred_lrate2
Improved
log rate
-4
-5
-6
-7
-8
-9
40
50
60
70
80
agegrp

# Model 3: Categorical age and smoker as the predictor

- Comparing Model 1 and 2, change in deviance, $85.01 - 23.99 = 61.02$, follows a $\chi^2_3$ under the null that log death rate is constant for different decades of age increase. Highly significant.
- Model 2 still has a large deviance, thought it improves a lot than Model 1.
- Add `smoker` to Model 2.
- $\log E(death_i) = \log(personyear_i) + \beta_0 + \beta_1 age_{1i} + \beta_2 age_{2i} + \beta_3 age_{3i} + \beta_4 age_{4i} + \beta_5 smoker_i$
- Now $\beta_4$ represents the incremental log death rate for age group 75-84 versus baseline age group 35-44, adjusting for the smoking status.

```
. poisson death smoker i.age, exposure(personyr) nolog

Poisson regression                              Number of obs     =         10
                                                LR chi2(5)        =     922.93
                                                Prob > chi2       =     0.0000
Log likelihood = -33.600153                     Pseudo R2         =     0.9321
------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   .3545356   .1073741     3.30   0.001     .1440862     .564985
             |
         age |
          1  |   1.484007   .1951034     7.61   0.000     1.101611    1.866403
          2  |   2.627505   .1837273    14.30   0.000     2.267406    2.987604
          3  |   3.350493   .1847992    18.13   0.000     2.988293    3.712693
          4  |   3.700096   .1922195    19.25   0.000     3.323353     4.07684
             |
       _cons |  -7.919326   .1917618   -41.30   0.000    -8.295172   -7.543479
 ln(personyr)|          1  (exposure)
------------------------------------------------------------------------------
. poisgof

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

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

Even better!


-4
-5
-6
-7
-8
-9
log rate
40
50
60
70
80
agegrp
lrate, smoker == 0
lrate, smoker == 1
pred_lrate3, smoker == 0
pred_lrate3, smoker == 1

# Model 4: Add interaction term

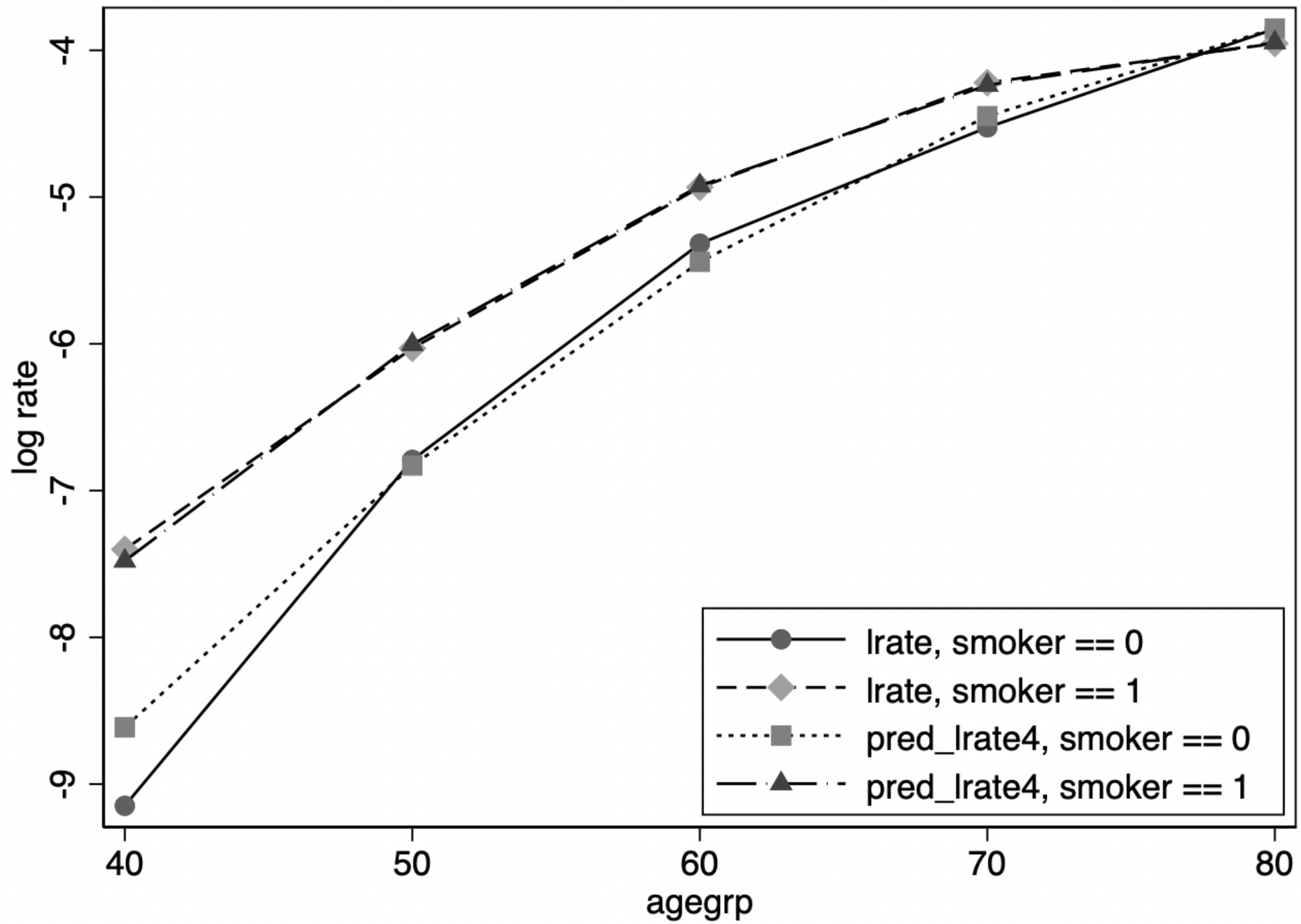
- Comparing Model 3 and 2, change in deviance, $23.00 - 12.13 = 11.87$, follows a $\chi_1^2$ under the null that smoker is not an important predictor. Highly significant.
- Model 3 still has a large deviance, thought adding smoker improves a lot than Model 2.
- The incremental log death rate across gender is shrinking as age increases. Consider adding interaction between `smoker` and `age`.
- In Stata, `xi` is a command to factorize the categorical variable and expand the variable's interaction.
- Note that this is a full model (10 parameters for 10 groups of counts).

```
. xi: poisson death i.age*smoker, exposure(personyr) nolog
  i.age             _Iage_0-4           (naturally coded; _Iage_0 omitted)
  i.age*smoker      _IageXsmoke_#       (coded as above)

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

  ------------------------------------------------------------------------------
         death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
       _Iage_1 |   2.357367   .7637625     3.09   0.002     .8604198    3.854314
       _Iage_2 |   3.830163    .731925     5.23   0.000     2.395616    5.264709
       _Iage_3 |   4.622656    .731925     6.32   0.000      3.18811    6.057203
       _Iage_4 |   5.294359   .7295601     7.26   0.000     3.864448    6.724271
        smoker |   1.746873   .7288689     2.40   0.017     .3183163     3.17543
  _IageXsmoke_1 |  -.9866227   .7900624    -1.25   0.212    -2.535117    .5618712
  _IageXsmoke_2 |  -1.362809   .7561868    -1.80   0.072    -2.844908    .1192903
  _IageXsmoke_3 |   -1.44229   .7565319    -1.91   0.057    -2.925065    .0404855
  _IageXsmoke_4 |  -1.846991   .7571736    -2.44   0.015    -3.331024   -.3629584
         _cons |  -9.147933   .7071067   -12.94   0.000    -10.53384   -7.762029
   ln(personyr) |          1  (exposure)
  ------------------------------------------------------------------------------
. poisgof

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

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

# Model 5: Add squared age

- There are strong interaction effects.
- Maybe treating `age` as continuous and considering the age-by-smoker interaction, as well as the quadratic age effect (`agesq`)? The effect of age on death is non-linear. Since association may not be linear.
- $\log E(death_i) =$
  $\log(personyear_i) + \beta_0 + \beta_1 age_i + \beta_2 smoker_i + \beta_3 agesq_i + \beta_4 sa_i$

```
. gen agesq = age*age
. gen sa = smoker*age
. poisson death age smoker agesq sa, exp(personyr) nolog
  Poisson regression                              Number of obs   =         10
                                                  LR chi2(4)      =     933.43
                                                  Prob > chi2     =     0.0000
  Log likelihood = -28.351655                     Pseudo R2       =     0.9427
  ------------------------------------------------------------------------------
         death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
           age |   1.981125   .1602452    12.36   0.000      1.66705     2.2952
        smoker |   1.133424   .2807705     4.04   0.000     .5831238    1.683724
         agesq |  -.1976765   .0273674    -7.22   0.000    -.2513157   -.1440374
            sa |  -.3075481   .0970411    -3.17   0.002    -.4977452   -.1173509
         _cons |  -8.612961   .2917237   -29.52   0.000    -9.184729   -8.041193
   ln(personyr) |          1  (exposure)
  ------------------------------------------------------------------------------
. poisgof

  Deviance goodness-of-fit  =   1.63544
  Prob > chi2(5)            =    0.8969

  Pearson goodness-of-fit   =  1.550251
  Prob > chi2(5)            =    0.9072
```



log rate
-4
-5
-6
-7
-8
-9
40
50
60
70
80
agegrp
lrate, smoker == 0
lrate, smoker == 1
pred_lrate4, smoker == 0
pred_lrate4, smoker == 1

# Model summary

| Model | Predictor | df | Deviance | AIC |
|---|---|---|---|---|
| 1 | age | 2 | 85 | 144 |
| 2 | ai.age | 5 | 24 | 89 |
| 3 | ai.age + smoker | 6 | 12 | 79 |
| 4 | ai.age x smoker | 10 | 0 | 75 |
| 5 Best. | aage + agesq + smoker | 5 | 2 | 67 |

Note: Similar as in logistic regression, the deviance and Pearson goodness-of-fit tests only apply to grouped (Poisson) data. You need other tests to assess the goodness-of-fit of a model for ungrouped data. Difference in deviance (LR) test can still be used to compare nested models.

- You may also directly output incident rate ratio (IRR), $\exp(\beta_1), \ldots, \exp(\beta_k)$, by using the option "`irr`".

```
. poisson death age smoker agesq sa, irr exp(personyr) nolog

Poisson regression                              Number of obs     =         10
                                                LR chi2(4)        =     933.43
                                                Prob > chi2       =     0.0000
Log likelihood = -28.351655                     Pseudo R2         =     0.9427

------------------------------------------------------------------------------
       death |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.250897   1.161922    12.36   0.000     5.296522    9.926422
      smoker |   3.106274   .8721498     4.04   0.000     1.791626    5.385573
       agesq |   .8206353   .0224587    -7.22   0.000     .7777768    .8658554
          sa |   .7352475   .0713493    -3.17   0.002     .6078998    .8892731
       _cons |   .0001817    .000053   -29.52   0.000     .0001026    .0003219
ln(personyr) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

- The IRR for smoker is 3.106274, and $\log(3.106274) = 1.133424$ from the previous page.

## Poisson Regression

* When to use Poisson Regression?
• count / rate variable.
• Y or t : response

- Poisson regression models a Poisson-distributed count variable as the response.
- It models the nature log of the expected count as a linear combination of predictors (uses a log link function).
- The equal mean and variance assumption should always be checked when using a Poisson model. This can be done by goodness-of-fit tests and by examining whether the variance is close to mean. An alternative model to handle *over-dispersion* in count is the Negative Binomial model.
- In a Poisson regression, the offset term is used to model event rate, with exposure in person-time.
- Coefficients in Poisson regression are on the log(count) scale and have a multiplicative effect on the event rate.
- Hypothesis testing and model comparison can be done similarly as in a logistic regression.