Lecture 15: Poisson assumptions, offsets, and relative risk

Ani Manichaikul
amanicha@jhsph.edu

10 May 2007

Poisson regression models

- Log-linear model for mean rate:

$$log(\lambda_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  where p is the number of predictors (or covariates) in the model
- Random component:

$$Y_i | \mathbf{X_i} \sim \text{Poisson}(\lambda_i)$$

- Here, $\lambda_i = E(Y_i|X_i) = Var(Y_i|X_i)$

Exponentiating Poisson regression models

- Exponentiating gives us a model for the rate parameter, or expected counts:

$$\lambda_i = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

- For Poisson random variables, $\text{Mean}(Y_i) = \lambda_i$, so our log-linear model provides a prediction for the expected value of Y_i

Interpretation of the parameters β_j

- e^{β_j} = Rate ratio for a 1 unit increase in X_j, i.e. rate ratio for $X_j + 1$ compared to X_j, with other covariates held constant
- $e^{\Delta\beta_j}$ = Rate ratio for a Δ unit increase in X_j, i.e. rate ratio for $X_j + \Delta$ compared to X_j, with other covariates held constant
- e^{β_0} = Baseline rate value, i.e. rate for an observation with all $\mathbf{X}$'s equal to zero

Estimation

- Estimates of the β's are obtained using maximum likelihood (or maximum quasi-likelihood)
- Estimates of the variances are usually obtained by either:
  - Maximum likelihood: assumes variance $= \lambda_i$ (the poisson rate parameter) for each unique combination of predictors
  - Quasi-likelihood estimation: an extension of maximum likelihood, in which we can multiply the Poisson variance by a scale factor to allow for over/under dispersion compared to a Poisson distribution; more flexible modelling strategy which allows variances to differ from the expected values

Why model on the log scale?

- Our systematic portion of the model allows linear combinations of the covariates:

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Since we have no restrictions on the predictors $X_1, \ldots, X_p$, the predicted values can take any values on the real line: $(-\infty, +\infty)$
- But our outcome variable Y_i consists of counts, so the expected value of Y_i has the restriction:

$$\lambda_i \in [0, +\infty)$$

- After taking a log transform, we get:

$$\log(\lambda_i) \in \log\{[0, +\infty)\} = [\log\{0\}, \log\{+\infty\}) = (-\infty, +\infty)$$

which is just what we wanted

Modelling log outcomes

- After the log transform, in Poisson regression, we are modelling the log-expected count
- Our baseline coefficient β_0 will be interpreted as the log-expected count (or rate) in the baseline group, with all covariates set to zero
- Other coefficients will be interpreted as:
  - differences in log-expected counts
  - since $\log(\frac{a}{b}) = \log(a) - \log(b)$, we can also interpret them as the log ratio of expected counts (or log rate ratios)

Assumptions for Poisson regression I

- Just as with linear and logistic regression, we have the assumptions:
  - L: log transformed outcomes are **linearly** related to the predictor variables; hence the name log-linear regression can be used interchangeably with Poisson regression
  - I: outcomes are **independent** given covariates; if we know any outcome(s), that does not give us additional information about other outcomes beyond what is known from the model
- For Poisson regression, our distributional assumption is specified as $Y_i|\mathbf{X_i} \sim Poisson(\lambda_i)$

Assumptions for Poisson regression II

- The Poisson distribution assumption is actually quite strong and difficult to satisfy
- Recalling that for Poisson random variables: $\lambda_i = E(Y_i|X_i) = Var(Y_i|X_i)$, a possible diagnostic idea is to compare sample means and variance across similar levels of the covariates X

More on Poisson distribution assumptions I

We can actually state the assumptions underlying the Poisson distribution model more specifically:

1. Within any (extremely) small interval of space (or time) on which we are observing counts, Δt:
   - Pr(observe 1 event) $\approx \lambda \Delta t$
   - Pr(observe $>$ 1 event) $= o(\delta t)$, which means:

$$\lim_{\Delta t \to 0} \frac{\Pr(\text{observe} > 1 \text{ event})}{\Delta t} = 0$$

More on Poisson distribution assumptions II

2. The rate parameter λ is the same across all intervals: we can call this assumption a "homogeneity" assumption
3. Independent intervals: probability of observing an event in any particular interval does not depend on whether we observed event(s) in any other interval – we can think of independence as a "memorylessness" property

More on Poisson distribution assumptions III

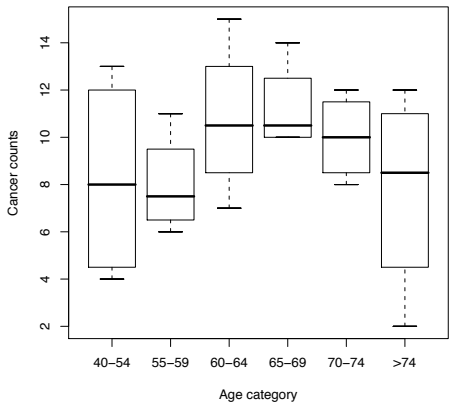
- The Poisson distribution certainly does not apply to any set of counts that we might observe
- It can be tricky to check the Poisson distributional assumptions
- We will need to think critically before applying these models
- For continuous covariates, it may be useful to group into quantiles and check estimated rates within grouped levels of predictors as as preliminary check of the model assumptions

Example: Danish Lung Cancer counts I

- Cases of lung cancer were counted in four Danish cities between 1968 and 1971 inclusive
- We have 24 observations on each of 4 variables:
  - Cases: the number of lung cancer cases
  - Pop: the population of each age group in each city
  - Age: the categorical age group; one of $40-54$, $55-59$, $60-64$, $65-74$ or >74
  - City: the city; one of Fredericia, Horsens, Kolding, or Vejle
- Questions of interest: How does the expected number of lung cancer counts vary by age?

Some plots to get started

Boxplots of observed counts versus age category



Model A: account for age only I

$$\begin{aligned}\log(\lambda_i) &= \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ &+ \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i)\end{aligned}$$

- We are fitting a model with indicators for each of the age categories
- Baseline is the group aged 40-54
- I(Age55-59) is an indicator of having age 55-59; it is equal to 1 for those of age 55-59 and 0 otherwise
- I(Age60-64) is an indicator of having age 60-64; it is equal to 1 for those of age 60-64 and 0 otherwise
- etc...

Model A: account for age only II

```
> summary(out.age <- glm(Cases~Age, family=poisson))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.11021    0.17408  12.122   <2e-16 ***
Age55-59    -0.03077    0.24810  -0.124    0.901
Age60-64     0.26469    0.23143   1.144    0.253
Age65-69     0.31015    0.22918   1.353    0.176
Age70-74     0.19237    0.23516   0.818    0.413
Age>74      -0.06252    0.25012  -0.250    0.803
```

Model A: account for age only III

$$\begin{aligned}\log(\lambda_i) &= 2.11 - 0.03I(\text{Age55-59}) + 0.265I(\text{Age60-64}) \\ &+ 0.310I(\text{Age65-69}) + 0.192I(\text{Age70-74}) - 0.06I(\text{Age}>74)\end{aligned}$$

- We interpret $\hat{\beta}_0 = 2.11$ as the log expected count of cancer cases among individuals aged 40-54
- We interpret $\hat{\beta}_0 + \hat{\beta}_1 = 2.08$ as the log expected count of cancer cases among individuals aged 55-59
- We interpret $\hat{\beta}_1 = -0.03$ as the difference in log expected count of cancer cases comparing the 55-59 age group to the 40-54 age group; We can also interpret $\hat{\beta}_1$ as a log relative rate

Model A: account for age only IV

$$\begin{aligned}\log(\lambda_i) &= 2.11 - 0.03I(\text{Age55-59}) + 0.265I(\text{Age60-64}) \\ &+ 0.310I(\text{Age65-69}) + 0.192I(\text{Age70-74}) - 0.06I(\text{Age}>74)\end{aligned}$$

- We interpret $\exp\{\hat{\beta}_0\} = 8.24$ as the expected count of cancer cases among individuals aged 40-54
- We interpret $\exp\{\hat{\beta}_0 + \hat{\beta}_1\} = 8.00$ as the expected count of cancer cases among individuals aged 55-59
- We interpret $\exp\{\hat{\beta}_1\} = 0.97$ as the ratio of expected counts comparing the 55-59 age group to the 40-54 age group; We can also interpret $\exp\{\hat{\beta}_1\}$ as a relative rate

Model A: account for age only V

Confidence intervals for all age coefficients contain 0... is there any association between cancer cases and age?

```
> confint(out.age)
                 2.5 %    97.5 %
(Intercept)  1.7484013 2.4330352
Age55-59    -0.5200788 0.4573101
Age60-64    -0.1863264 0.7248187
Age65-69    -0.1357451 0.7664976
Age70-74    -0.2671916 0.6587925
Age>74      -0.5565061 0.4289354
```


Model A: account for age only VI

Let's perform a likelihood ratio test to look at the global hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

versus the alternative hypothesis:

$$H_a : \text{at least one of the } \beta_i\text{'s is not 0, for } i \in 1, \ldots 5$$

logLik(Age model) = -59.57
logLik(intercept only model) = -62.04

Model A: account for age only VII

Test statistic:

$$\begin{aligned} \text{TS} &= \text{-2(logLik(intercept only model) - logLik(Age model))} \\ &= 4.95 \sim \chi^2_5 \text{ under the null hypothesis} \end{aligned}$$

Critical value for the hypothesis test at level $\alpha = 0.05$:

$$\chi^2_{5,1-0.05} = 11.07$$

Fail to reject the null hypothesis

Model A: account for age only VIII

Conclusions:

- Based on the Poisson model of cancer case counts as a function of Age, we noted a generally increasing number of cases with increasing age
- The trend was not monotonically increasing with age
- Not a statistically significant result

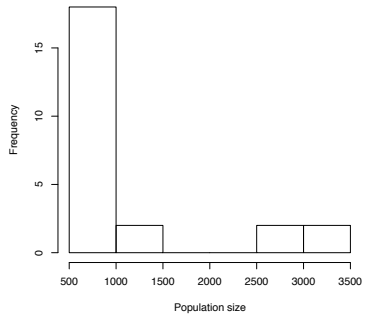
What about accounting for population size? I

- So far we modelled the observed counts of cancer cases as Poisson counts
- The population size from each of these counts was drawn is also known
- Can we improve our analysis?

- Each city and age group has a different population size
- If we model expected counts without accounting for population size, we may just be picking up effects of population distribution by age
- Accounting for population sizes can refine our analysis

What about accounting for population size? II

- The distribution of population sizes appears bimodal
- The population sizes range from a minimum of 509 to a maximum of 3142 people



What about accounting for population size? III

- Even within each age group, the population counts vary a bit
- For example, for ages 40-54, the four cities have population sizes 3059, 2879, 3142 and 2520
- For the greater than 74 age group, the four cities have populations 605, 782, 659 and 619
- So far, we have modeled expected counts for each population group, within the 4 year period of time:
  - Rate = Counts / 4 years

What about accounting for population size? IV

- It may be more interesting to know the rate per person, per 4 year period of time:
  - $\text{Rate} = \frac{\text{Counts/Population size}}{\text{4 years}} = \frac{\text{Counts}}{\text{4 person-years}}$
- Even better, we can get the rate per person year as:
  - $\text{Rate} = \frac{\text{Counts/(4 Population size)}}{\text{4 years}} = \frac{\text{Counts}}{\text{person-years}}$
- Here, $4 \cdot$ Population size is equal to the number of person-years that we observed to obtain each count
- We can think of the person years at the denominator to be used to calculate the cancer rate per person, per year
- If we prefer the cancer rate per 100 person-years, we can just multiply the rate per person-year by 100

Example: HIV infection rate

- Suppose we are told that 100,000 new cases of HIV were reported in the world, during the past three years
- What is the incidence rate of HIV?
- We can calculate incidence as:

$$\begin{aligned} & \frac{\text{Number of new cases}}{\text{Number of people observed} \cdot \text{Amount of time observed}} \\ = & \frac{100,000 \text{ cases}}{\text{6,000,000,000 people in the world} \cdot \text{3 years of observation}} \\ = & \ 5.55 \times 10^{-6} \text{ cases / person-year} \\ = & \ 5.55 \text{ cases/ 1,000,000 person-years} \end{aligned}$$

Based on this incidence rate, we could say that each year, there are about 5.55 new cases of HIV per 1,000,000 people per year

Example: Person years for Danish cancer cases in the 40-54 age group I

- Based on our regression model for counts on age, we calculated the coefficient $\hat{\beta}_0 = 2.11$, which we interpret as the log expected count in the baseline group of individuals age 40-54, during our 4 year period of observation from 1968-1971
- Exponentiating, we find $exp(2.11) = 8.25$ is the expected number of cancer cases per four year period of time
- The total population of individuals aged 40-54 in this study is 11,600

Example: Person years for Danish cancer cases in the 40-54 age group II

- So we can calculate: $\frac{8.25 \text{ counts in 4 years}}{11{,}600 \text{ people}} = 7.11 \times 10^{-4}$ is the rate of cancer cases in the population of individuals aged 40-54, per year
- Multiplying by 10000, we could get a more interpretable number, for instance:

$$\begin{aligned} & \frac{8.25 \text{ counts in 4 years}}{11{,}600 \text{ people}} \cdot \frac{10{,}000 \text{ people}}{10{,}000 \text{ people}} \\ = \; & 7.11 \text{ cases of cancer per 10,000 person-years} \end{aligned}$$

- If we observed 10,000 people aged 40-54 in a given year, we would expect to observe 7.11 cases of cancer among them

Modelling Danish cancer cases with an offset I

- So far, we have written a model for the expected number of counts over the 4 year period of observation
- However, if we know that the total populations generating our counts differ substantially, it does not make sense to write a log-linear model to consider expected counts direct
- What we really want is to consider, the rate per person year

$$r_i = \frac{\lambda_i}{\text{Pop}_i} = \frac{E(\text{count}_i)}{\text{Pop}_i}$$

  and model that by a log-linear model
- Based on this model, we can still say:

$$Y_i \sim \text{Poisson}(\lambda_i) = \text{Poisson}(r_i \cdot \text{Pop}_i)$$

Modelling Danish cancer cases with an offset II

On a log scale, our model will be:

$$\begin{aligned} \log(\frac{\lambda_i}{\text{Pop}_i}) &= \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ &+ \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i) \end{aligned}$$

Exponentiating, we get:

$$\begin{aligned} \frac{\lambda_i}{\text{Pop}_i} &= exp\{\beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ &+ \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i)\} \end{aligned}$$

- Since all counts were restricted to the same 4 year period of time from 1968 - 1971, the λ_i values are rates "per 4 years"
- Divide the λ_is by the population size that yielded each count to get rates "per 4 person-years"

Modelling Danish cancer cases with an offset III

- Now that we have a model of rates per 4 person-years, we can divide by 4 to get rate per person-year
- We can then multiply by 10,000 to get rates per 10,000 person-years (maybe easier to interpret than person-years)
- Dividing by 4 and then multiplying by 10,000 is the same as multiplying by 2500:

$$\begin{aligned} \frac{\lambda_i}{\text{Pop}_i/2500} &= \exp\{\beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ &+ \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i)\} \end{aligned}$$

Modelling Danish cancer cases with an offset IV

Finally, we should take a log-transform to get back our log-linear model:

$$\begin{aligned} & \log\left(\frac{\lambda_i}{\text{Pop}_i/2500}\right) \\ = \; & \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ + \; & \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i) \end{aligned}$$

Further, we can move the population denominator to the other size of the equation:

$$\begin{aligned} \log(\lambda_i) \; = \; & \log(\text{Pop}_i/2500) + \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ + \; & \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}{>}74_i) \end{aligned}$$

Modelling Danish cancer cases with an offset V

- Here, we call the amount $\log(\text{Pop}_i/2500)$ the offset
- Using the offset is just a way of accounting for population sizes, which could vary by age, region, etc.
- The term "offset" is jargon for predictor terms whose β coefficient is forced to be $+1$
- It is easy to model the offset in R , and most other statistical packages
- Using an offset gives us a convenient way to model rates per person-years, instead of just modeling the raw counts

Fitting a log-linear model with offset in R I

We can use the same command as usual to fit our model with an offset, and the only extra work is in specifying the offset value:

```
> summary(out.offset <- glm(Cases~Age
       + offset(log(Pop/2500)), family=poisson, data=lung))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9618     0.1741  11.270  < 2e-16 ***
Age55-59      1.0823     0.2481   4.363 1.29e-05 ***
Age60-64      1.5017     0.2314   6.489 8.66e-11 ***
Age65-69      1.7503     0.2292   7.637 2.22e-14 ***
Age70-74      1.8472     0.2352   7.855 4.00e-15 ***
Age>74        1.4083     0.2501   5.630 1.80e-08 ***
```


Fitting a log-linear model with offset in R II

We can also get confidence intervals for the regression parameters as before:

```
> confint(out.offset)
                2.5 %   97.5 %
(Intercept) 1.5999812 2.284615
Age55-59    0.5930347 1.570424
Age60-64    1.0506566 1.961802
Age65-69    1.3043866 2.206629
Age70-74    1.3876585 2.313643
Age>74      0.9142950 1.899736
```

All of our parameters are statistically significant at level α =5%

Interpretation of the model with offset I

- After including the offset in our model, we need to change our regression coefficient interpretations a bit
- We should think of the outcome as $\log(\lambda_i) - \text{offset}_i$
- In this case, λ_i was the expected number of cases observed in a particular age group and city, within a 4 year period of time
- Our offset was $\log(\text{Pop}_i/2500)$
- So, we should think of the outcome as log rate per 10,000 person years

Interpretation of the model with offset II

- β_0 is the log rate of cancer cases per 10,000 person years in the baseline age group of 40-54
- β_1 is the log relative rate of cancer cases per 10,000 person years comparing the age group 55-59 to the baseline age group 40-54
- β_2 is the log relative rate of cancer cases per 10,000 person years comparing the age group 60-64 to the baseline age group 40-54

Statistical significance of the age model I

- Before we put the offset in our model, none of our regression coefficients were statistically significant
  $\Rightarrow$ without the offset, there was no statistically significant difference in the expected counts per year at age group compared to the baseline 40-54 group
- After including the offset, we're looking for differences in the expected counts per person-year, across age groups
- This is a different question...

Statistical significance of the age model II

Let's perform a likelihood ratio test to look at the global hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

versus the alternative hypothesis:

$$H_a : \text{at least one of the } \beta_i\text{'s is not 0, for } i \in 1, \ldots 5$$

logLik(Age model with offset) = -62.35
logLik(intercept model, with offset) = -113.148

Statistical significance of the age model III

Test statistic:

$$\begin{aligned} \text{TS} &= \text{-2(logLik(intercept only, with offset)} \\ &- \text{logLik(Age model with offset))} \\ &= 101.6 \sim \chi^2_5 \text{ under the null hypothesis} \end{aligned}$$

Critical value for the hypothesis test at level $\alpha = 0.05$:

$$\chi^2_{5,1-0.05} = 11.07$$

Reject the null hypothesis, p-value = `pchisq(101.6, df=5, lower.tail=F)` $= 2.4 \times 10^{-20}$

Statistical significance of the age model IV

- You can see that the hypothesis test with and without the offset terms are looking at very different things
- Without the offset, we failed to reject, while with the offset we got a p-value less than 0.001
- Without the offset, we were thinking about differences in expected cases of cancer across the age groups
  $\Rightarrow$ there may not have been much of a difference because the slightly lower population in higher age groups was counterbalanced by high rate of cancer with increasing age
- With the offset, we were focused on differenced in cancer rates **per person-years** across the age groups
  $\Rightarrow$ the offset lets us compare the rate of acquiring lung cancer among those who are alive, i. e. taking into account the number of people within in cohort of interest

Predictions using the offset model I

$$\begin{aligned} \log(\lambda_i) &= \log(\text{Pop}_i/2500) + \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ &+ \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age>74}_i) \end{aligned}$$

Predicted log expected count of cancer cases from 1968-1971 among 40-54 year olds:

$$\begin{aligned} \log(\lambda_i) &= \log(\frac{\text{number of 40-54 year olds}}{2500}) + \hat{\beta}_0 \\ &= \log(\frac{11600}{2500}) + 1.96 = 3.49 \end{aligned}$$

Predicted log rate of cancer per 10,000 person years among 40-54 year olds:

$$\log(\lambda_i) = \hat{\beta}_0 = 1.96$$

Predictions using the offset model II

$$\begin{aligned} \log(\lambda_i) &= \log(\text{Pop}_i/2500) + \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ &+ \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age>74}_i) \end{aligned}$$

Predicted expected count of cancer cases from 1968-1971 among 40-54 year olds:

$$\begin{aligned} \lambda_i &= exp(\log(\frac{\text{number of 40-54 year olds}}{2500}) + \hat{\beta}_0) \\ &= exp(3.49) = 32.9 \end{aligned}$$

Predicted rate of cancer per 10,000 person years among 40-54 year olds:

$$\lambda_i = exp(\hat{\beta}_0) = 7.09$$

Based on this model, we predict 7.09 new cases of lung cancer per 10,000 40-54 year olds in Denmark, per year

Predictions using the offset model III

$$\begin{aligned}\log(\lambda_i) &= \log(\text{Pop}_i/2500) + \beta_0 + \beta_1 I(\text{Age55-59}_i) + \beta_2 I(\text{Age60-64}_i) \\ &+ \beta_3 I(\text{Age65-69}_i) + \beta_4 I(\text{Age70-74}_i) + \beta_5 I(\text{Age}>74_i)\end{aligned}$$

Predicted log expected count of cancer cases from 1968-1971 among 55-59 year olds:

$$\begin{aligned}\log(\lambda_i) &= \log(\frac{\text{number of 55-59 year olds}}{2500}) + \hat{\beta}_0 + \hat{\beta}_1 \\ &= \log(\frac{3811}{2500}) + 1.96 + 1.08 = 3.46\end{aligned}$$

Predicted log rate of cancer per 10,000 person years among 55-59 year olds:

$$\log(\lambda_i) = \hat{\beta}_0\hat{\beta}_1 = 1.96 + 1.08 = 3.04$$

Predictions using the offset model IV

Predicted expected count of cancer cases from 1968-1971 among 55-99 year olds:

$$\begin{aligned}\lambda_i &= exp(log(\frac{\text{number of 55-59 year olds}}{2500}) + \hat{\beta}_0 + \hat{\beta}_1) \\ &= exp(3.46) = 31.8\end{aligned}$$

Predicted rate of cancer per 10,000 person years among 55-59 year olds:

$$\lambda_i) = exp(\hat{\beta}_0 + \hat{\beta}_1) = exp(1.96 + 1.08) = exp(3.04) = 20.9$$

Based on this model, we predict 20.9 new cases of lung cancer per 10,000 55-59 year olds in Denmark, per year

Predictions using the offset model V

- We predicted an incidence rate of 7.09 cases per 10,000 per year among 40-54 year olds, and a rate of 20.9 new cases per 10,000 person years for 55-59 years olds
  $\Rightarrow$ The relative rate per 10,000 person years comparing 55-59 years olds to 40-54 years olds is 20.9 / 7.09 $\approx$ 2.94
- We could have gotten the same answer by taking $exp(\hat{\beta}_1) = exp(1.08) = 2.94$
- $exp(\beta_1)$ is the relative rate of cancer cases per 10,000 person years comparing 55-59 years olds to 40-54 years olds
- β_1 is the log relative rate

Confidence intervals for relative rates I

We can get confidence intervals for relative rates by exponentiating the confidence intervals for regression conefficients:

```
> exp(confint(out.offset))
                2.5 %    97.5 %
(Intercept) 4.952939  9.821906
Age55-59    1.809471  4.808685
Age60-64    2.859528  7.112129
Age65-69    3.685428  9.085042
Age70-74    4.005460 10.111189
Age>74      2.495016  6.684133
```

Confidence intervals for relative rates II

- There is an increasing trend in relative rates compared to the baseline 40-54 year old group as age increases
- The only exception for this trend is for the Age>74 group
- The model fit matches what we know from biology: the risk of cancer does increase with age, but trails of for the oldest individuals perhaps because
  - those surviving to age 74 and beyond have genes which protect against cancer
  - cell growth slows down at older ages, slowing the growth of tumors

Some additional notes about offsets

- The purpose of an offset is the change the denominator or units of a rate
- Often, the model without an offset does not make much sense, and likely fails our Poisson assumptions
- We should always try to use an offset if we suspect that the underlying population sizes differ for each of the observed counts
- Typically the offset will take value $\log(N)$ where N is the sample size, or the number of person-years
- If the underlying population sizes are not available, we just have to do our best – but be careful about applying the Poisson model

Poisson regression for cohort studies I

A bit of background

- We can think of relative rate as the ratio comparing rates in two groups: $\frac{R_1}{R_2}$ where R_1 is the rate for group 1, and R_2 is the rate for group 2
- We can extend rates for counts to binary outcomes, i.e. relative risk is $\frac{p_1}{p_2}$ where p_1 is the probability of the event in group 1 compared to the probability p_2 for group 2
- Relative risk is a convenient way of comparing disease risk across different strata of the population, such as male vs. female, young vs. old, etc.

Poisson regression for cohort studies II

- Generally, it is more convenient to look at relative risk $(\frac{p_1}{p_2})$ than odds ratio $(\frac{p_1/(1-p_1)}{p_2/(1-p_2)})$
- So why did get into odds ratios in the first place?
- Original motivation: Case control studies sample a fixed number of cases and controls, so we have no way to estimate the overall probability of disease in any group, can only estimate odds
- However, in a cohort study, we follow a fixed group of people for a long period of time, and observed the rate at which disease cases come about
- In cohort studies, we can estimate the actual rates of disease

Poisson regression for cohort studies III

- We can use log-linear models to look at data from cohort studies, and then our coefficients are interpreted as risk, and relative risks
- Warning: should only use log-linear models for binary outcomes that are rare
- The risk of disease is supposed to be a number between 0 and 1, while traditional log-linear models were designed for any positive rates (from 0 to ∞)

Poisson regression for cohort studies IV

- One trick people sometimes use is to perform log-linear models, with a restriction to rates between 0 and 1 (in R , specify `family=binomial(link=log)`), though this trick can also lead to trouble...
- Another way to get relative risks would have been to perform logistic regression, get odds, and transform those to probabilities – a valid strategy for cohort studies (but not case control)

Summary I

- Log-linear models can be a good way to approach count data
- If population sizes or denominators are available, it's a good idea to include an offset
- Log-linear models can also be useful in analyzing binary data from cohort studies, but with care

Summary II

- Poisson regression models fall under the family of generalized linear models
- Many tools we used for linear and logistic regression apply here as well
- For hypothesis testing: Wald test and likelihood ratio tests
- For modelling:
  - Adjustment for confounding by additive covariates
  - Effect modification modelled using interaction terms
  - Polynomial and spline terms for curves or bends