Nested case-control studies

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Design of Epidemiological Studies

So far

- We have estimates rate ratios, comparing those with high and low alcohol intake
- In this analysis we divided the follow-up into two wide time bands; age 50 to 60, and age 60 to 70
- We assumed that the rate is constant within these time bands

Unrealistic assumption

- The assumption of constant rates between age 50 and 60, and between age 60 and 70, is probably unrealistic
- If the assumption is (very) wrong, then our results may be (very) wrong as well
- How can we relax this assumption?

Outline

Poisson regression with narrow time bands

Cox regression

The nested case-control study

Outline

Poisson regression with narrow time bands

Cox regression

The nested case-control study

Narrower time bands

- Let's see what happens if we split follow-up into
  - 5 year time bands
  - 1 year time bands

  and assume that the rate is constant within these

Outcomes and person-years with 5 year time bands

	alc.binary	timeband	pyears	n	event
1	0	1	2484.153	497	1
2	1	1	2515.000	503	0
3	0	2	2471.376	496	6
4	1	2	2500.808	503	9
5	0	3	2403.437	490	20
6	1	3	2374.405	494	47
7	0	4	2276.196	470	31
8	1	4	2058.258	447	75

Poisson regression with 5 year time bands

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-9.0498	1.0056	-9.000	< 2e-16	***
alc.binary	0.8749	0.1578	5.546	2.92e-08	***
timeband2	2.7136	1.0322	2.629	0.00856	**
timeband3	4.2550	1.0068	4.226	2.38e-05	***
timeband4	4.8296	1.0041	4.810	1.51e-06	***

- The estimated alcohol effect is similar to before and fairly precise; small standard error and statistically significant
- The estimated effect of each time band increases with time, and is also fairly precise

Outcomes and person-years with 1 year time bands

	alc.binary	timeband	pyears	n	event
1	0	1	497.0000	497	0
2	1	1	503.0000	503	0
3	0	2	497.0000	497	0
4	1	2	503.0000	503	0
5	0	3	497.0000	497	0
6	1	3	503.0000	503	0
7	0	4	497.0000	497	0
8	1	4	503.0000	503	0
9	0	5	496.1533	497	1
10	1	5	503.0000	503	0

Data with 1 year time bands, cont'd

	alc.binary	timeband	pyears	n	event
11	0	6	496.0000	496	0
12	1	6	503.0000	503	0
13	0	7	496.0000	496	0
14	1	7	502.2998	503	1
15	0	8	495.0308	496	2
16	1	8	501.0363	502	1
17	0	9	493.0130	494	2
18	1	9	498.6010	501	5
19	0	10	491.3320	492	2
20	1	10	495.8713	496	2

Data with 1 year time bands, cont'd

	alc.binary	timeband	pyears	n	event
21	0	11	488.7974	490	3
22	1	11	492.4695	494	4
23	0	12	484.8234	487	5
24	1	12	487.9795	490	7
25	0	13	480.9281	482	2
26	1	13	474.9227	483	13
27	0	14	477.3498	480	6
28	1	14	465.1198	470	11
29	0	15	471.5387	474	4
30	1	15	453.9131	459	12

Data with 1 year time bands, cont'd

	alc.binary	timeband	pyears	n	event
31	0	16	467.5907	470	4
32	1	16	442.9370	447	7
33	0	17	462.3463	466	6
34	1	17	431.4997	440	15
35	0	18	457.2019	460	6
36	1	18	415.3450	425	21
37	0	19	449.2820	454	11
38	1	19	393.7769	404	20
39	0	20	439.7748	443	4
40	1	20	374.6995	384	12

Poisson regression with 1 year time bands

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.182e+01  8.017e+02  -0.027    0.978
alc.binary   8.800e-01  1.578e-01   5.578 2.44e-08 ***
timeband2   -1.196e-07  1.134e+03   0.000    1.000
timeband3   -1.196e-07  1.134e+03   0.000    1.000
timeband4   -1.196e-07  1.134e+03   0.000    1.000
timeband5    1.437e+01  8.017e+02   0.018    0.986
timeband6   -4.227e-04  1.134e+03   0.000    1.000
timeband7    1.437e+01  8.017e+02   0.018    0.986
timeband8    1.548e+01  8.017e+02   0.019    0.985
timeband9    1.633e+01  8.017e+02   0.020    0.984
timeband10   1.577e+01  8.017e+02   0.020    0.984
timeband11   1.634e+01  8.017e+02   0.020    0.984
timeband12   1.689e+01  8.017e+02   0.021    0.983
timeband13   1.713e+01  8.017e+02   0.021    0.983
timeband14   1.727e+01  8.017e+02   0.022    0.983
timeband15   1.723e+01  8.017e+02   0.021    0.983
timeband16   1.688e+01  8.017e+02   0.021    0.983
timeband17   1.755e+01  8.017e+02   0.022    0.983
timeband18   1.783e+01  8.017e+02   0.022    0.982
timeband19   1.801e+01  8.017e+02   0.022    0.982
timeband20   1.739e+01  8.017e+02   0.022    0.983
```

- The estimated alcohol effect is similar to before and fairly precise; small standard error and statistically significant
- However, the estimated time band effects are highly uncertain; large standard errors and non-significant
- *Why does this happen?*

The no-interaction assumption

- The model assumes no interaction between alcohol intake and time
  - same effect of alcohol across time bands
  - same effect of each time band across alcohol levels
- To estimate the alcohol effect the model can pool information across all 20 time bands
  - large gain in precision
- To estimate the effect of each time band, the model can only pool information across the 2 alcohol levels
  - small gain in precision

A trade-off

- If we use very wide time bands, then we make unrealistic model assumptions
- If we use very narrow time bands, then we obtain uncertain estimates of the time band effects
- A reasonable compromise is often to use 5 year time bands

What is the research question?

- Often (but not always!) we are only interested in the exposure effect
- If so, it doesn't matter that the estimated time band effects are uncertain
- Then it is best to use narrow time bands, to make the model as realistic as possible

Outline

Poisson regression with narrow time bands

Cox regression

The nested case-control study

Regression models for cohort studies with time-to-event data

- Poisson regression is probably the most common model among epidemiologists to estimate rates and rate ratios
- However there are many alternative models, all with different pros and cons
- One common alternative to Poisson regression is **Cox regression**

Cox regression

- Cox regression allows the rates to vary continuously in time
  - no assumption about constant rate within time bands
- Advantage: less modeling assumptions than Poisson regression
- Disadvantage: cannot estimate the effect of time
  - but are you really interested in this?

Poisson regression vs Cox regression

- Poisson regression:

```
> fit <- glm(formula=death~alc.binary+timeband+
  offset(log(tdiff/365.25)), family="poisson",
  data=cohort.split)
> summary(fit)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.0498     1.0056  -9.000  < 2e-16 ***
alc.binary    0.8749     0.1578   5.546 2.92e-08 ***
timeband2     2.7136     1.0322   2.629  0.00856 **
timeband3     4.2550     1.0068   4.226 2.38e-05 ***
timeband4     4.8296     1.0041   4.810 1.51e-06 ***
```

- Cox regression:

```
> fit <- coxph(formula=Surv(time, death)~alc.binary,
  data=cohort)
> summary(fit)
              coef exp(coef) se(coef)     z Pr(>|z|)
alc.binary 0.8790    2.4086   0.1578 5.572 2.52e-08 ***
```

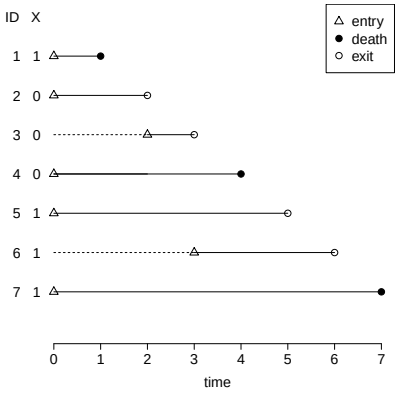
From Cox regression to nested case-control studies

- The theory behind Cox regression gives intuition for how the nested case-control study works
- Let's look closer into this

Risk sets

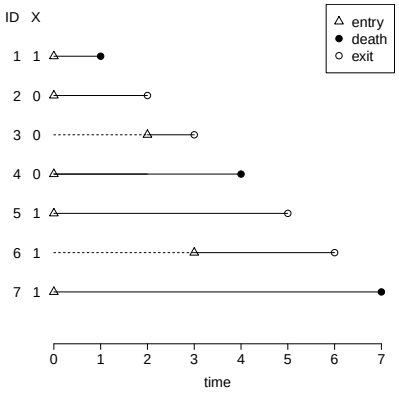
- The theory for Cox regression relies heavily on the notion of **risk sets**
- For each time point t where there is at least one observed outcome, there is a risk set
- The risk set at t contains those subjects who are observed at t

Example



- *Which are the risk sets?*

Solution



- Observed outcomes at times 1, 4 and 7
- Risk set at time 1: subjects 1, 2, 4, 5, 7
- Risk set at time 4: subjects 4, 5, 6, 7
- Risk set at time 7: subject 7

Original data vs risk set data

- Original data:

ID	X	entry	exit	D
1	1	0	1	1
2	0	0	2	0
3	0	2	3	0
4	0	0	4	1
5	1	0	5	0
6	1	3	6	0
7	1	0	7	1

- Risk set data:

ID	X	Time	Fail	Set
1	1	1	1	1
2	0	1	0	1
4	0	1	0	1
5	1	1	0	1
7	1	1	0	1
4	0	4	1	2
5	1	4	0	2
6	1	4	0	2
7	1	4	0	2
7	1	7	1	3

How Cox regression works

- Cox regression first transforms the original data into risk set data
- For each risk set it compares the exposure distribution between
  - those in the risk set who have the outcome at the time point when the risk set is formed, and
  - those in the risk set who don't have the outcome at the time point when the risk set is formed
- An imbalance in the exposure distribution gives a weight to the estimated log rate ratio
  - larger proportion of exposed among those who have the outcome weights the estimate upwards
  - larger proportion of exposed among those who don't have the outcome weights the estimate downwards

Example

ID	X	Time	Fail	Set
1	1	1	1	1
2	0	1	0	1
4	0	1	0	1
5	1	1	0	1
7	1	1	0	1
4	0	4	1	2
5	1	4	0	2
6	1	4	0	2
7	1	4	0	2
7	1	7	1	3

- In risk set 1:
  - 1/1=100% of those in the risk set with an outcome are exposed
  - 2/4=50% of those in the risk set without an outcome are exposed
- Larger proportion of exposed among those who have the outcome; upward weight to the estimated log rate ratio

Example, cont'd

ID	X	Time	Fail	Set
1	1	1	1	1
2	0	1	0	1
4	0	1	0	1
5	1	1	0	1
7	1	1	0	1
4	0	4	1	2
5	1	4	0	2
6	1	4	0	2
7	1	4	0	2
7	1	7	1	3

- In risk set 2:
  - 0/1=0% of those in the risk set with an outcome are exposed
  - 3/3=100% of those in the risk set without an outcome are exposed
- Larger proportion of exposed among those who have the outcome; downward weight to the estimated log rate ratio

Example, cont'd

ID	X	Time	Fail	Set
1	1	1	1	1
2	0	1	0	1
4	0	1	0	1
5	1	1	0	1
7	1	1	0	1
4	0	4	1	2
5	1	4	0	2
6	1	4	0	2
7	1	4	0	2
7	1	7	1	3

- In risk set 3:
  - 1/1=100% of those in the risk set with an outcome are exposed
  - No subject in the risk set without an outcome; nothing to compare with
- No weight to the estimated rate ratio

Example, cont'd

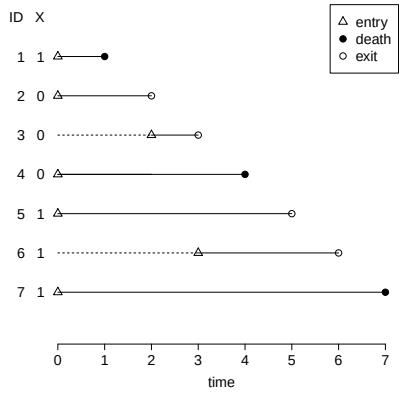
ID	X	Time	Fail	Set
1	1	1	1	1
2	0	1	0	1
4	0	1	0	1
5	1	1	0	1
7	1	1	0	1
4	0	4	1	2
5	1	4	0	2
6	1	4	0	2
7	1	4	0	2
7	1	7	1	3

- *Do the risk set data remind you of any other design we have talked about?*

Cox regression as a matched case-control study

- Cox regression can be thought of as treating the cohort study as a matched case-control study
- The risk sets are the matched sets
- The cases in a set are the subjects who have the outcome at the time when the risk set is formed
- The controls in a set are the subjects who don't have the outcome at the time when the risk set is formed
- The number of cases and controls vary across sets
- Within each set, time is fixed
- **The controls are matched to the cases on the underlying time scale**

The choice of baseline

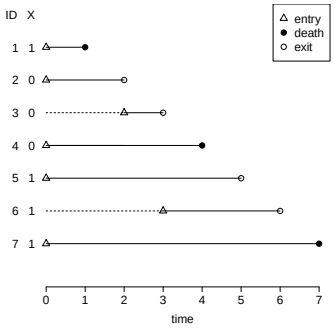


- *What would happen to the risk sets if we re-define baseline as enrollment into the study?*

Solution

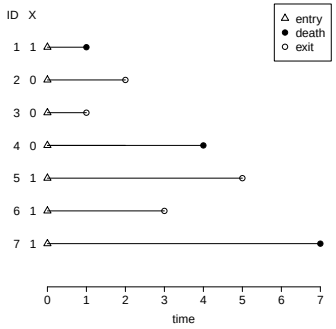
- The baseline determines how subjects are 'aligned' to each other, which defines the risk sets

- Original baseline



- Subject 6 is in the risk set at time 4

- New baseline



- Subject 6 is not in the risk set at time 4

Conditional logistic regression?

- Previously, we have used conditional logistic regression to analyze matched case-control data
- Let's see what happens if we
  - rearrange the cohort data into matched case-control data, as outlined above, and
  - use conditional logistic regression to analyze these matched case-control data

Rearrangement of data

```
> library(Epi)
> cohort.risksets <- ccwc(exit=time, fail=death,
  controls=nrow(cohort), data=cohort,
  include=list(sex, alc.binary))
> cohort.risksets <-
  cohort.risksets[order(cohort.risksets$Time), ]
```

Data; first risk set from baseline

```
> cohort.risksets[1:12, ]
   Set Map Time Fail sex alc.binary
1  176 934 1517    1   0          0
2  176 415 1517    0   1          1
3  176 813 1517    0   0          0
4  176  48 1517    0   0          0
5  176 845 1517    0   1          1
6  176 672 1517    0   1          1
7  176 854 1517    0   1          1
8  176 100 1517    0   1          1
9  176 519 1517    0   0          0
10 176 252 1517    0   1          1
11 176 382 1517    0   1          0
12 176 284 1517    0   0          0
> sum(cohort.risksets$Set==176 &
  cohort.risksets$Fail==0)
[1] 999
```

- 999 controls

Data; second risk set from baseline

```
> cohort.risksets[1001:1012, ]
     Set Map Time Fail sex alc.binary
1001  27 133 2301    1   1          1
1002  27 479 2301    0   0          1
1003  27 463 2301    0   1          1
1004  27 714 2301    0   0          0
1005  27 349 2301    0   1          0
1006  27 202 2301    0   1          1
1007  27 184 2301    0   0          0
1008  27 270 2301    0   0          0
1009  27 344 2301    0   0          0
1010  27 278 2301    0   1          1
1011  27 691 2301    0   1          1
1012  27 299 2301    0   0          1
> sum(cohort.risksets$Set==27 &
  cohort.risksets$Fail==0)
[1] 998
```

- 998 controls

But wait a second..

- Doesn't Cox regression and conditional logistic regression estimate different parameters?
  - the estimated parameter in Cox regression is a (log) rate ratio
  - the estimated parameter in conditional logistic regression is a (log) odds ratio
- Why would these be equal?

Cox regression vs. conditional logistic regression

- Cox regression:

```
> fit <- coxph(formula=Surv(time, death)~alc.binary,
  data=cohort)
> summary(fit)
              coef exp(coef) se(coef)     z Pr(>|z|)
alc.binary 0.8790    2.4086   0.1578 5.572 2.52e-08 ***
```

- Conditional logistic regression:

```
> fit <- clogit(Fail~alc.binary+strata(Set),
  data=cohort.risksets)
> summary(fit)
              coef exp(coef) se(coef)     z Pr(>|z|)
alc.binary 0.8790    2.4086   0.1578 5.572 2.52e-08 ***
```

- Identical estimates and standard errors

The odds ratio as a rate ratio

- The conditional logistic regression estimates the odds ratio

$$\frac{\text{odds of death at } t \text{ for the exposed, given survival up to } t}{\text{odds of death at } t \text{ for the unexposed, given survival up to } t}$$

- The model assumes that this odds ratio is constant across all time points t
- At any t, the risk of death is small
  - even if the risk is large over a longer period
- Hence, the odds ratio is approximately equal to the risk ratio

$$\frac{\text{risk of death at } t \text{ for the exposed, given survival up to } t}{\text{risk of death at } t \text{ for the unexposed, given survival up to } t}$$

...which by definition is the rate ratio at t

Outline

Poisson regression with narrow time bands

Cox regression

The nested case-control study

The need for subsampling

- Sometimes, we want to restrict data collection or analysis to a subsample of the cohort
  - it may be difficult or expensive to collect data on the exposure and confounders
  - the full cohort may be very large, and take a long time to analyze
- Viewing the cohort study as a matched case-control study leads to a natural framework for subsampling

Unnecessarily many controls per case

- In matched case-control studies with only one case per set, the benefit of each additional control decreases rapidly

- In the motivating example, we have 999 controls in the first set, 998 in the second set etc
  - many more than needed for reasonable precision in the estimates

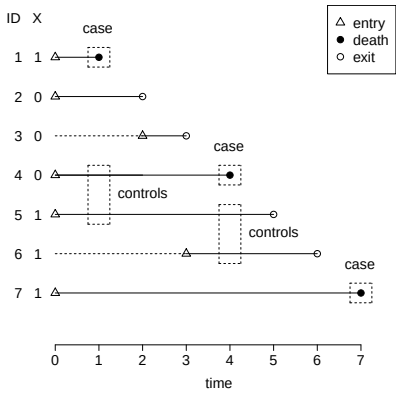
The nested case-control study

- The **nested case-control study** aims to reduce the size of the cohort, by sampling a small number of controls at random from each risk set
- This sampling scheme is referred to as **incidence density sampling**
- Incidence density sampling increases statistical uncertainty, but usually not very much

Exposure distribution in the nested case-control study

- In each risk set, Cox regression compares the exposure distribution between cases and controls
- Incidence density sampling takes controls at random, from each risk set
- Thus, the exposure distribution is similar among the controls in the nested case-control sample as among the controls in the full cohort, in each risk set
- Incidence density sampling gives an unbiased estimate of the exposure effect

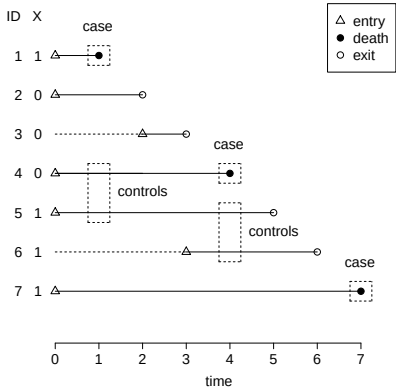
Example



ID	*X*	Time	Fail	Set
1	1	1	1	1
~~2~~	~~0~~	~~1~~	~~0~~	~~1~~
4	0	1	0	1
5	1	1	0	1
~~7~~	~~1~~	~~1~~	~~0~~	~~1~~
4	0	4	1	2
5	1	4	0	2
6	1	4	0	2
~~7~~	~~1~~	~~4~~	~~0~~	~~2~~
7	1	7	1	3

- *Compare the exposure distribution among the controls in the full cohort and in the nested case-control sample, in each risk set*

Solution



ID	*X*	Time	Fail	Set
1	1	1	1	1
~~2~~	~~0~~	~~1~~	~~0~~	~~1~~
4	0	1	0	1
5	1	1	0	1
~~7~~	~~1~~	~~1~~	~~0~~	~~1~~
4	0	4	1	2
5	1	4	0	2
6	1	4	0	2
~~7~~	~~1~~	~~4~~	~~0~~	~~2~~
7	1	7	1	3

Risk set	controls in cohort	controls in nested case-control
1	50% exposed	50% exposed
2	100% exposed	100% exposed

Incidence density sampling

```
> nested.case.control <- ccwc(exit=time,
  fail=death, controls=10, data=cohort,
  include=list(sex, alc.binary))
> nested.case.control <-
  nested.case.control[order(nested.case.control$Time), ]
```

Cohort study vs. nested case-control study with 10 controls per case

- Cohort study, Cox regression:

```
> fit <- coxph(formula=Surv(time, death)~
  alc.binary, data=cohort)
> summary(fit)
             coef exp(coef) se(coef)     z Pr(>|z|)
alc.binary 0.8790    2.4086   0.1578 5.572 2.52e-08 ***
```

- Nested case-control study, conditional logistic regression:

```
> fit <- clogit(Fail~alc.binary+strata(Set),
  data=nested.case.control)
> summary(fit)
             coef exp(coef) se(coef)     z Pr(>|z|)
alc.binary 0.8532    2.3471   0.1637 5.212 1.87e-07 ***
```

- Similar estimates and standard errors

Similarity with cohort studies

- The nested case-control study arises by subsampling from a cohort
- Thus, the nested case-control study has a baseline and an underlying time scale
- If there is truncation/censoring in the cohort, then there is truncation/censoring in the nested case-control sample
- In this sense, the nested case-control study is similar to the cohort study

Similarity with matched case-control studies

- In nested case-control studies, the controls are matched to the cases on the underlying time scale
- In this sense, the nested case-control study is similar to the matched case-control study

Further matching

- In nested case-control studies it is common to additionally match on potential confounders
- This is done by sampling cases and controls with the same confounder levels from each risk set
- Just like in 'ordinary' matched case-control studies, we must control for the matching variables in the analysis
  - automatically done with conditional logistic regression

Summary

- Poisson regression assumes that the rate is constant within time bands
  - wider time bands gives a more unrealistic model
  - narrower time bands gives more uncertain estimates of the time band effects
- Cox regression allows the rates to vary continuously in time
- In effect, Cox regression treats the cohort study as matched (on time) case-control study
- With incidence density sampling we sample a small number of controls at random from each risk set
- The resulting study is called a nested case-control study
- Conditional logistic regression analysis of nested case-control data usually gives similar results as Cox regression analysis of full cohort data