Biostatistics III: Survival analysis for epidemiologists in R

Alexander Ploner
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
http://www.biostat3.net/

4–13 November, 2024

https://doctoralcourses.application.ki.se/fubasextern/info?kurs=C8F2992, course code C8F2992

Topics for Day 2

- Estimands and estimators
- Hazards and rates
- Time scales
- Modelling rates, using Poisson regression
- Interactions and parameterisation
- Confounding by time

Estimands, estimators and estimates

- An estimand is what we (conceptually) want to calculate – which often relates to our research question. An example is the proportion of individuals who are alive at five years after study entry.
- An estimator is a calculation process (e.g. a formula or an algorithm) to calculate an estimate. An example is the Kaplan-Meier estimator for survival (that is, the formula).
- An estimate is the resulting calculation from applying an estimator to some data. An example would be the "estimated" five-year survival from a Kaplan-Meier estimator from a particular study with follow-up to a specific date.
- We often need to consider how to interpret an estimator for a given study design. For example, an odds ratio estimate using a conditional logistic regression estimator from a nested case-control study with incidence density sampling can be interpreted as a hazard ratio – which may be the estimand of interest.

Estimands and estimators

Estimand	Estimator	Notes
Survival	Kaplan-Meier	Non-parametric
	Poisson regression	Awkward post-estimation
	Cox model + Breslow	Proportional hazards
	Flexible parametric model	
Cumulative hazard	Nelson-Aalen	Non-parametric
Hazard	Nelson-Aalen + Kernel density	Smoothed
Rate	count/(person-time)	Poisson distribution
	Poisson regression	
Rate ratio	Poisson regression	
Hazard ratio	Cox model	Non-parametric baseline
	Flexible parametric model	
Time-dependent rate ratio	Poisson model	
Time-dependent hazard ratio	Cox model	Inflexible implementations
	Flexible parametric model	

Hazard rates and the hazard function, $\lambda(t)$ I

- In contrast to the survival function, which describes the probability of *not* failing before time t, the hazard function focuses on the failure rate at time t among those individuals who are alive at time t. So, the survival function is formally defined for a random time variable T by

$$S(t) = \Pr(T > t) = 1 - F(t). \tag{1}$$

  where $F(t)$ is the failure proportion (aka the cumulative distribution function).

- The hazard function is formally defined for a random time variable T by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{2}$$

- The hazard function shows how the hazard rate varies over time.
- The hazard function, $\lambda(t)$, is the instantaneous event rate at time t, conditional on survival up to time t.

Hazard rates and the hazard function, $\lambda(t)$ II

- From Equation 2, one can see that $\lambda(t)\Delta t$ may be viewed as the 'approximate' probability of an individual who is alive at time t experiencing the event in the next small time interval Δt.
- The units are events per unit time.
- Note that the hazard is a rate, not a probability, so $\lambda(t)$ can take on any value between zero and infinity, as opposed to $S(t)$ which is restricted to the interval $[0, 1]$.
- A lower value for $\lambda(t)$ implies a higher value for $S(t)$ and *vice versa*.

Hazard rates and the hazard function, $\lambda(t)$ III

- There is a close relationship between the hazard function and survival. First, let the hazard λ be constant. Then survival $S(t)$ between time 0 and time t is approximately $1 - \lambda t$. We can improve on this calculation by breaking t into n segments and calculate the probability of surviving each of the segments; letting n become large, we have that

$$\begin{aligned} S(t) &= \lim_{n\to\infty} \prod_{i=1}^{n} \left(1 - \lambda\frac{t}{n}\right) \\ &= \lim_{n\to\infty} \prod_{i=1}^{n} \exp\left(-\lambda\frac{t}{n}\right) \\ &= \lim_{n\to\infty} \exp\left(-\sum_{i=1}^{n} \lambda\frac{t}{n}\right) \\ &= \exp(-\lambda t) \end{aligned}$$

This is the survival function for an exponential distribution.

Hazard rates and the hazard function, $\lambda(t)$ IV

- More generally, for a time-varying hazard, we have that

$$
\begin{aligned}
S(t) &= \lim_{n\to\infty} \prod_{i=1}^{n} \left(1 - \lambda\left(t\frac{i}{n}\right)\frac{t}{n}\right) \\
&= \lim_{n\to\infty} \prod_{i=1}^{n} \exp\left(-\lambda\left(t\frac{i}{n}\right)\frac{t}{n}\right) \\
&= \exp\left(-\lim_{n\to\infty} \sum_{i=1}^{n} \lambda\left(t\frac{i}{n}\right)\frac{t}{n}\right) \\
&= \exp\left(-\int_0^t \lambda(u)du\right) \qquad (3)
\end{aligned}
$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is the area under the hazard function between 0 and t (the cumulative hazard).

Hazard rates and the hazard function, $\lambda(t)$ V

- ☡ As noted earlier, the hazard function is the negative change in log survival, such that

$$-\frac{d}{dt}\log(S(t)) = \lambda(t)$$

and the hazard function is the rate of decline in survival, such that

$$-\frac{dS(t)}{dt}/S(t) = \lambda(t)$$

- ☡ Some statisticians will explain that the hazard is the instantaneous rate, where the rate $r(s,t)$ between times s and t can be formally defined as

$$r(s,t) = \frac{\text{Expected count}}{\text{Expected person-time}} = \frac{\int_s^t S(u)\lambda(u)du}{\int_s^t S(u)du}$$

and, if s and t are close, then $r(s,t) \approx \lambda\left(\frac{s+t}{2}\right)$.

Choice of time scale

- There are several time scales along which rates might vary. These differ from one another only in the choice of *time origin*, the point at which time is zero.
- Consider the following questions?
  - What is the time?
  - How old are you?
  - For how long have you lived at your current address?
- What is the time origin for each?
- In which units did you specify time? Could different units have been used?
- Time progresses in the same manner but, in answering these questions, we have applied a different time origin and used different units.

A sample of 35 patients diagnosed with colon carcinoma during 1985–94; followed-up until the end of 1995

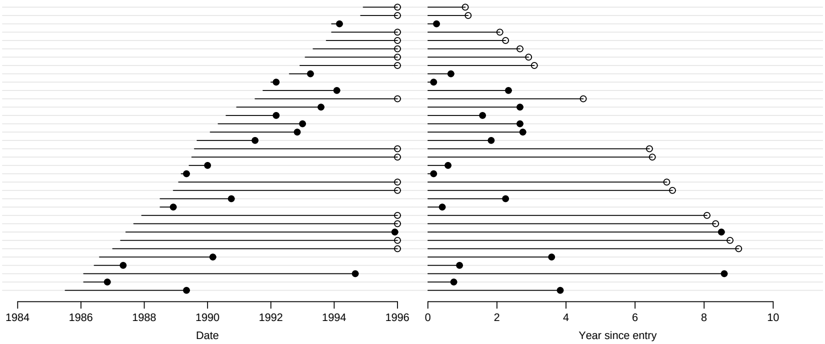


Figure: Calendar time (left) and time from entry in years (right)

Common time scales in epidemiology

Origin	Time scale
Birth	Age
A fixed date	Calendar time
First exposure	Time exposed
Entry into study	Time in study
Disease onset	Time since onset
Diagnosis	Time since diagnosis
Start of treatment	Time on treatment

- In many of the methods used in survival analysis, effects are adjusted for the underlying time scale. Choice of time scale therefore has important implications.
- On many time scales, subjects do not enter follow-up at the time origin, $t = 0$.
- To deal with these issues, the `Surv` function allows for both the entry and exit times to be specified prior to the event indicator.

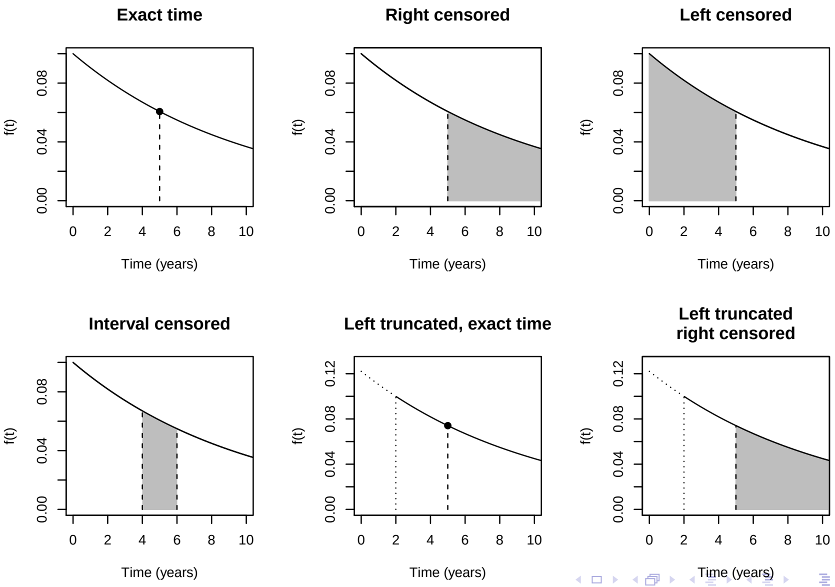
Censoring and truncation I

- With right censoring, the most common form of censoring in medical studies, we know that the event has not occurred during follow-up, but we are unable to follow-up the patient further. We know only that the true survival time of the patient is greater than a given value.
- Less common is left-censoring, where we know the event has occurred prior to the time of observation but we don't know exactly when.
- Interval censoring occurs when we know that the event has occurred between two time points but don't know the exact date (e.g. HIV infection between two test dates, or cancer between two screens).
- Standard methods for survival analysis assume that all censored data are right censored and we will assume that this is the case.
- Special methods are required for analysing left censored and interval censored data, which will not be covered in this course.
- Censoring, in general, refers to the situation where we can identify the individuals in our study but we do not have precise information on the event time for all individuals (we know only that it is in some interval).

Censoring and truncation II

- A second feature of survival studies, often confused with censoring, is truncation.
- Truncation refers to the situation where certain subjects are screened such that the investigator is not aware of their existence.
- Left truncated data occurs when we only observe the individual if they are event free after a certain follow-up time. For example, late entry to the study or using age as the primary time scale.
- The distribution is conditional on entry at the left truncation time.
- Left truncation will change who is in the risk set at different times and the person-time observed.
- Methods for left truncated data are available for:
  1. Hazard estimation
  2. Poisson regression
  3. Cox regression
  4. Generalised survival models
- Right truncated data occurs when only individuals who experience the event of interest are included in the study. Special methods of analysis are required for their analysis (see Klein & Moeschberger [1]).

Graphical representation of censoring and truncation



Informative right-censoring I

- To make it possible for statistical analysis we make the crucial assumption that, conditional on the values of any explanatory variables, censoring is unrelated to prognosis (the probable course and outcome of the disease).
- The statistical methods used for survival analysis assume that the prognosis for an individual censored at time t will be no different from those individuals who were alive at time t and were under follow-up past time t.
- One way to think of this is that, conditional on the values of any explanatory variables, the individuals censored at time t should be a random sample of the individuals at risk at time t.
- This is known as noninformative censoring. Under this assumption, there is no need to distinguish between the different reasons for right-censoring.
- When withdrawal from follow-up is associated with prognosis, this is known as informative censoring and standard methods of analysis will result in biased estimates.
- Common methods for controlling for informative censoring are to stratify or condition on those explanatory factors on which censoring depends.

Informative right-censoring II

- Censoring due to termination of the study, or accidental death, are usually uninformative, but careful consideration must be given to other forms of censoring.
- Determining whether or not censoring is informative is not a statistical issue — it must be made based on subject matter knowledge.

The diet data I

- For the diet data
  - date of entry = doe
  - date of exit = dox
  - event indicator = chd

R code

```
Surv(tstop,event) # right censored
Surv(tstart,tstop,event) # left truncated and right censored
```

Defines time to event for many survival analysis procedures.
To use time since entry as the time scale:

R code and output

```
> scale <- 365.24
> with(diet, Surv((dox-doe)/scale, chd))
[1] 16.7916986+ 19.9594787+ 19.9594787+ 15.3953565+  1.4949075
```

The diet data II

- Each individual enters the study (becomes ‘at risk’) at follow-up time zero (doe is the calendar time origin).
- By dividing by 365.24 we are scaling the time unit from days to years.
- The event is defined by chd.
- To use attained age as the time scale we specify

R code and output

```
> with(diet, Surv((doe-dob)/scale, (dox - dob)/scale, chd))
[1] (49.38944,66.18114+] (47.49754,67.45701+] (46.46534,66.42482+]
```

- Individuals enter the study at doe (as before) but the time origin is now the date of birth.
- To use calendar time as the time scale we specify (using continuous years rather than Dates)

R code

```
Surv(yoe, yox, chd)
```

The diet data III

- Equivalently, we could define the origin as 1 January 1900 and then use Dates:

R code and output

```
> origin <- as.Date("1900-01-01")
> with(diet, Surv((doe-origin)/scale, (dox - origin)/scale, chd))
[1] (60.12485,76.91655+] (56.95707,76.91655+] (56.95707,76.91655+]
```

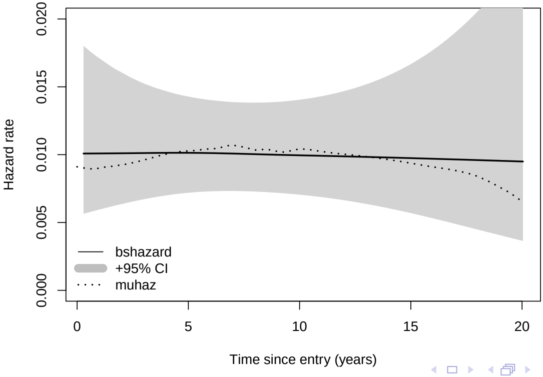
Hazard smoothing

- For a non-parametric estimator of survival, we can calculate the steps in the cumulative hazard function. If we smooth those steps, we can estimate the hazard function – much like we do for density estimation.
- The main challenge with the smoothing is choosing the bandwidth, or the amount of data, to include in the smoother.
- Advice: do not over-interpret the estimated smooth hazard functions.
- The biostat3 package provides a Surv interface to the muhaz function from the muhaz package. For right censored and possibly left truncated data, we suggest using the bshazard function from the bshazard package.

CHD rate with time-since-entry as the time scale

R code

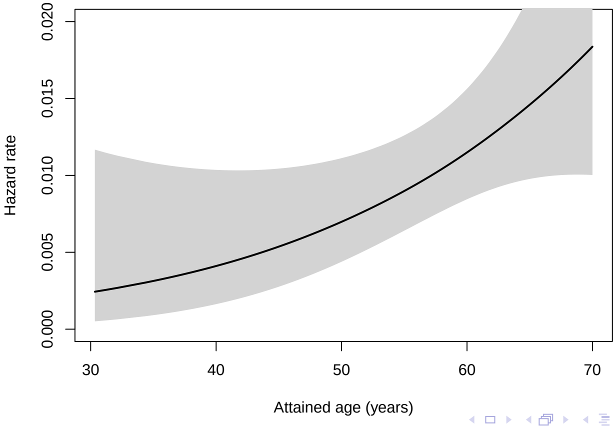
```
plot(bshazard(Surv((dox-doe)/365.24, chd) ~ 1, data=diet),
     ylim=c(0,0.02), xlab="Time since entry (years)")
lines(muhaz2(Surv((dox-doe)/365.24, chd) ~ 1, data=diet),
      lty=3, lwd=2)
legend("bottomleft",legend=c("bshazard","+95% CI","muhaz"), lty=1:3,
       col=c("black","grey","black"), lwd=c(1,10,2), bty="n")
```



CHD rate with attained age as the time scale

R code

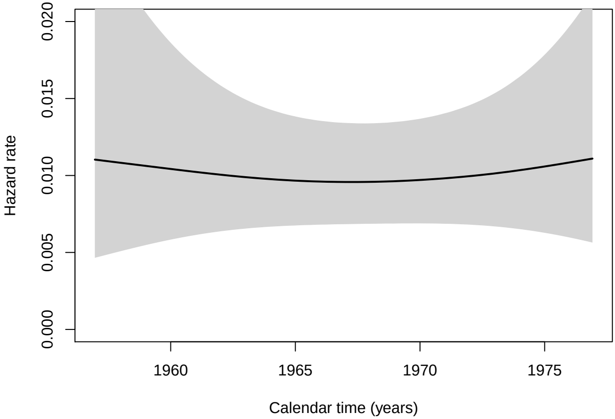
```
plot(bshazard(Surv((doe-dob)/365.24, (dox-dob)/365.24, chd) ~ 1,
              data=diet),
     ylim=c(0,0.02), xlab="Attained age (years)")
```



CHD rate with calendar time as the time scale

R code

```
plot(bshazard(Surv(year(doe), year(dox), chd) ~ 1, data=diet),
     ylim=c(0,0.02),xlab="Calendar time (years)")
```



R digression: formulas and factors I

- Let the linear predictor for the right-hand-side of a formula be represented by η (pronounced "eta"), where $\eta = X\beta$ for design matrix X and parameters β
- Example data (similar to the diet data-frame):

```
df = data.frame(hieng=factor(c("low","low","low","high","high","high"),
                             levels=c("low","high")),
                job=factor(c("driver","conductor","bank",
                             "driver","conductor","bank"),
                           levels=c("driver","conductor","bank")),
                x=1:6)
```

Continuous var with intercept (ncol=2)

```
> model.matrix(~x,df)
  (Intercept) x
1           1 1
2           1 2
3           1 3
4           1 4
5           1 5
6           1 6
```

$$\eta = \beta_0 + \beta_1 x$$

Continuous var - intercept (ncol=1)

```
> model.matrix(~x-1,df)
  x
1 1
2 2
3 3
4 4
5 5
6 6
```

$$\eta = \beta_1 x$$

R digression: formulas and factors II

2-level factor with intercept (ncol=2)

```
> model.matrix(~hieng,df)
  (Intercept) hienghigh
1           1         0
2           1         0
3           1         0
4           1         1
5           1         1
6           1         1
```

$$\eta = \beta_0 + \beta_1 I(hieng = high)$$

3-level factor with intercept (ncol=3)

```
> model.matrix(~job,df)
  (Intercept) jobconductor jobbank
1           1            0       0
2           1            1       0
3           1            0       1
4           1            0       0
5           1            1       0
6           1            0       1
```

$$\eta = \beta_0 + \beta_1 I(job = conductor) + \beta_2 I(job = bank)$$

2-level factor - intercept (ncol=2)

```
> model.matrix(~hieng-1,df)
  hienglow hienghigh
1        1         0
2        1         0
3        1         0
4        0         1
5        0         1
6        0         1
```

$$\eta = \beta_1 I(hieng = low) + \beta_2 I(hieng = high)$$

3-level factor - intercept (ncol=3)

```
> model.matrix(~job-1,df)
  jobdriver jobconductor jobbank
1         1            0       0
2         0            1       0
3         0            0       1
4         1            0       0
5         0            1       0
6         0            0       1
```

$$\eta = \beta_1 I(job = driver) + \beta_2 I(job = conductor) + \beta_3 I(job = bank)$$

R digression: formulas and factors III

Main effects model with hieng and job:

```
> model.matrix(~hieng+job,df)
  (Intercept) hienghigh jobconductor jobbank
1           1         0            0       0
2           1         0            1       0
3           1         0            0       1
4           1         1            0       0
5           1         1            1       0
6           1         1            0       1
```

$$\eta = \beta_0 + \beta_1 I(hieng = high) + \beta_2 I(job = conductor) + \beta_3 I(job = bank)$$

Main effect and interaction model with hieng and job:

```
> model.matrix(~hieng+job+hieng:job,df)
  (Intercept) hienghigh jobconductor jobbank hienghigh:jobconductor hienghigh:jobbank
1           1         0            0       0                      0                 0
2           1         0            1       0                      0                 0
3           1         0            0       1                      0                 0
4           1         1            0       0                      0                 0
5           1         1            1       0                      1                 0
6           1         1            0       1                      0                 1
```

$$\eta = \beta_0 + \beta_1 I(hieng = high) + \beta_2 I(job = conductor) + \beta_3 I(job = bank) + \beta_4 I(hieng = high \;\&\; job = conductor) + \beta_5 I(hieng = high \;\&\; job = bank)$$

R digression: formulas and factors IV

job main effect and interaction between hieng and job:

```
>   model.matrix(~job+hieng:job,df)
  (Intercept) jobconductor jobbank jobdriver:hienghigh jobconductor:hienghigh
1           1            0       0                   0                      0
2           1            1       0                   0                      0
3           1            0       1                   0                      0
4           1            0       0                   1                      0
5           1            1       0                   0                      1
6           1            0       1                   0                      0
  jobbank:hienghigh
1                 0
2                 0
3                 0
4                 0
5                 0
6                 1
```

$$\eta = \beta_0 + \beta_1 I(job = conductor) + \beta_2 I(job = bank) + \beta_3 I(hieng = high \,\&\, job = driver) + \beta_4 I(hieng = high \,\&\, job = conductor) + \beta_5 I(hieng = high \,\&\, job = bank)$$

R digression: formulas and factors V

hieng main effect and interaction between hieng and job:

```
> model.matrix(~hieng+hieng:job,df)
  (Intercept) hienghigh hienglow:jobconductor hienghigh:jobconductor hienglow:jobbank
1           1         0                     0                      0                0
2           1         0                     1                      0                0
3           1         0                     0                      0                1
4           1         1                     0                      0                0
5           1         1                     0                      1                0
6           1         1                     0                      0                0
  hienghigh:jobbank
1                 0
2                 0
3                 0
4                 0
5                 0
6                 1
```

$$\eta = \beta_0 + \beta_1 I(hieng = high) + \beta_2 I(hieng = low \;\&\; job = conductor) + \beta_3 I(hieng = high \;\&\; job = conductor) + \beta_4 I(hieng = low \;\&\; job = bank) + \beta_5 I(hieng = high \;\&\; job = bank)$$

R digression: formulas and factors VI

hieng and x main effects:

```
>   model.matrix(~hieng+x,df)
  (Intercept) hienghigh x
1           1         0 1
2           1         0 2
3           1         0 3
4           1         1 4
5           1         1 5
6           1         1 6
```

$$\eta = \beta_0 + \beta_1 I(hieng = high) + \beta_2 x$$

hieng and x main effects and interaction between hieng and x:

```
> model.matrix(~hieng+x+hieng:x,df)
>   model.matrix(~hieng+x+hieng:x,df)
  (Intercept) hienghigh x hienghigh:x
1           1         0 1           0
2           1         0 2           0
3           1         0 3           0
4           1         1 4           4
5           1         1 5           5
6           1         1 6           6
```

$$\eta = \beta_0 + \beta_1 I(hieng = high) + \beta_2 x + \beta_3 I(hieng = high)x$$

Estimating CHD rates according to energy intake I

- The new `biostat3::survRate` function tabulates the number of events and person-time at risk and calculates event rates.

R code

```
> survRate(Surv((dox-doe)/365.24/1000, chd) ~ hieng, data=diet)
           hieng    tstop event      rate    lower    upper
hieng=low    low 2.059487    28 13.595619 9.034190 19.64946
hieng=high  high 2.544308    18  7.074616 4.192866 11.18094
```

- The unit for the person-time `T` is 1000 person-years and the rates are per 1000 person-years.
- The rates represent the overall rates of CHD in each group during follow-up.

Estimating CHD rates according to energy intake II

- The incidence rate ratio (IRR) for individuals with a high compared to low energy intake is $7.1/13.6 = 0.52$.
- That is, without controlling for any possible confounding factors, we estimate that individuals with a high energy intake have a CHD risk that is approximately half that of individuals with a low energy intake.
- This is sometimes called a ‘crude estimate’; it is not adjusted for potential confounders.
- Is this a true effect? What important confounder might we need to consider?

A model for the incidence rate I

- When working with rates, we believe that effects are most likely to be multiplicative.
- That is, we believe that the rate in the high energy group (λ_1) is likely to be a multiple of the rate in the low energy group (λ_0). The multiplication factor is the incidence rate ratio, θ.

$$\lambda_1 = \lambda_0\theta, \text{ for example, } 7.1 = 13.6 \times 0.52$$

$$\text{IRR} = \frac{\lambda_1}{\lambda_0} = \theta, \text{ for example, } 0.52 = 7.1/13.6$$

A model for the incidence rate II

- If the explanatory variable X is equal to 1 for individuals with a high energy intake and 0 for individuals with a low energy intake then we can write

$$\lambda(X) = \lambda_0 \times \theta^X$$

- So for each increase of one unit in X the rate increases with a multiple of θ, i.e. the effects are multiplicative.
- That is,

$$\lambda = \lambda_0 \text{ when } X = 0$$

$$\lambda = \lambda_0\theta \text{ when } X = 1$$

- For instance, the rate λ_1 among the individuals with high energy intake is

$$\lambda_1 = \lambda(1) = \lambda_0 \times \theta^1 = 13.6 \times 0.52 = 7.1$$

A model for the incidence rate III

- In practice, it is more convenient to work on a logarithmic scale.

$$\begin{aligned} \lambda &= \lambda_0 \times \theta^X \\ &= \exp\left(\log\left(\lambda_0 \times \theta^X\right)\right) \\ &= \exp(\log(\lambda_0) + X\log(\theta)) \\ &= \exp(\beta_0 + \beta_1 X) \end{aligned}$$

  where $\beta_0 = \log(\lambda_0)$ and $\beta_1 = \log(\theta)$ is the log IRR.

- On the log scale, the effects are additive. For an increase of one unit in X, the log rate increases with an constant $\log(\theta)$, or β_1.
- $\log(\lambda) = \beta_0 + \beta_1 X$ is a Poisson regression model with one binary explanatory variable, X.
- Exercise: What are the estimates of β_0 and β_1?

A model for the incidence rate IV

- The estimate of β_0 is the log of the rate at baseline, log(13.6)
- The estimate of β_1 is the log of the IRR comparing group 1 to group 0, log(0.52)

Three regression models commonly applied in epidemiology

- Linear regression

$$\mu = \beta_0 + \beta_1 X$$

- Logistic regression

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

- Poisson regression

$$\log(\lambda) = \beta_0 + \beta_1 X$$

- In each case β_1 is the effect per unit of X, measured as a change in the mean (linear regression); the change in the log odds (logistic regression); the change in the log rate (Poisson regression).

The effect of high energy, using Poisson regression

hieng	X	D	Y	Rate per 1000
low	0	28	2059.4	13.60
high	1	18	2544.2	7.07

- If we assume a Poisson regression model

$$\begin{aligned}
\log(\lambda) &= \beta_0 + \beta_1 X \\
X = 0 : \log(28/2059.4) &= \beta_0 = -4.3 \\
X = 1 : \log(18/2544.2) &= \beta_0 + \beta_1 \\
\log\left(\frac{18/2544.2}{28/2059.4}\right) &= \beta_1 \\
-0.6532 &= \beta_1 = \log(IRR) \\
0.52 &= \exp(\beta_1) = IRR
\end{aligned}$$

Poisson regression model for rates

- We assume that the counts are Poisson distributed. If the count is Y and the person-time is T (assumed fixed), then for the above model we have that

$$\begin{aligned} \lambda = E\left(\frac{Y}{T}\right) &= \exp(\beta_0 + \beta_1 X) \\ \implies E(Y) &= \exp(\beta_0 + \beta_1 X)T \\ &= \exp(\beta_0 + \beta_1 X + \log(T)) \end{aligned}$$

  where $\log(T)$ is termed an offset, which is a component of the linear predictor that does not have a coefficient.

- Importantly, if we use the count as the outcome and the log of the person-time as an offset in a Poisson regression, then we are modelling rates.

Poisson regression in R I

R code and output

```
> library(broom) # tidy()
> diet <- transform(diet, y = as.numeric(dox-doe)/365.24)
> fit <- glm(chd ~ hieng + offset(log(y)), data=diet, family=poisson)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2980     0.1890 -22.744   <2e-16 ***
hienghigh    -0.6532     0.3021  -2.162   0.0306 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> broom::tidy(fit, conf.int=TRUE, exponentiate=TRUE)
# A tibble: 2 × 7
  term        estimate std.error statistic   p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)   0.0136     0.189    -22.7  1.63e-114  0.00916    0.0193
2 hienghigh     0.520      0.302     -2.16 3.06e-  2  0.283      0.933
```

Poisson regression in R II

- Poisson regression is one type of generalised linear model (GLM)
- The Poisson model is estimated using the method of maximum likelihood.
- Confidence intervals are constructed by assuming the estimated regression parameters are normally distributed.
- That is, confidence intervals are constructed on the log scale, as is standard for ratio measures, with an interval $(\exp(\hat{\beta} - 1.96 se), \exp(\hat{\beta} + 1.96 se))$.
- We see that the confidence limits for the IRR are simply the exponentiated limits of the log IRR.
- As such, the CI for the IRR is not symmetric around the point estimate.

What happened to the time scale?

- The Poisson model we just fitted did not take into account that rates may vary over follow-up time.
- We used time since entry as the time scale, but the rate we estimated was the ‘overall rate’ of CHD throughout the follow-up, i.e. simply all events of CHD divided by total person-time at risk.
- When we estimate the *overall rate*, we assume that the rates (13.6 per 1,000 person-years among low energy group, and 7.1 among high energy group) are constant throughout the follow-up time.
- We will get back to how to model rates which vary over time later. But first we will have a look at how to model main effects and interactions, in general, in Poisson regression.

Categorical exposures with more than two levels I

- The variable hieng has two levels
- The variable eng3, created below, has 3 levels.

R code and output

```
> diet <- transform(diet, eng3 = cut(energy, c(0,1500,2500,3000,4500,Inf),
                                     right=FALSE))
> survRate(Surv((dox-doe)/365.24/1000, chd) ~ eng3, data=diet)
                           tstop event      rate    lower    upper
eng3=[1.5e+03,2.5e+03) 0.9466597    16 16.901532 9.660686 27.44703
eng3=[2.5e+03,3e+03)   2.0173174    22 10.905572 6.834464 16.51117
eng3=[3e+03,4.5e+03)   1.6398177     8  4.878591 2.106229  9.61277
```

Categorical exposures with more than two levels II

- To include eng3 in regression functions we can use indicator variables (0/1 or FALSE/TRUE) for the 3 levels.

R code and output

```
> diet <- transform(diet,
                    X1 = energy<2500,
                    X2 = energy>=2500 & energy<3000,
                    X3 = energy>=3000 & energy<4500)
> tidy(glm(chd ~ X2 + X3 + offset(log(y)), data=diet, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)   0.0169     0.250    -16.3  6.64e-60  0.00991    0.0266
2 X2TRUE        0.645      0.329     -1.33 1.82e- 1  0.341      1.25
3 X3TRUE        0.289      0.433     -2.87 4.11e- 3  0.117      0.656
```

- The variable ($X1$) that indicates the category with the lowest energy intake is omitted, meaning this is the reference category.

Categorical exposures with more than two levels III

- In terms of the parameters

$$\begin{aligned}
\log(\lambda) &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 \\
\\
&= \beta_0 \qquad\qquad\qquad \text{(level 1)} \\
&= \beta_0 + \beta_2 \quad \text{(level 2)} \\
&= \beta_0 + \beta_3 \quad \text{(level 3)}
\end{aligned}$$

Automatic generation of indicators using factor variables I

- The baseline is, by default, the first level, but this can be changed to (say) the third level (3000–) with

R code and output

```
> transform(diet, eng3 = relevel(eng3,"[3e+03,4.5e+03)")) |>
      glm(formula=chd ~ eng3 + offset(log(y)), family=poisson) |>
      tidy(conf.int=TRUE, exponentiate=TRUE)
  term                   estimate std.error statistic  p.value conf.low conf.high
  <chr>                     <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)             0.00488     0.354    -15.1  3.18e-51  0.00223   0.00908
2 eng3[1.5e+03,2.5e+03)  3.46        0.433      2.87 4.11e- 3  1.52      8.54
3 eng3[2.5e+03,3e+03)    2.24        0.413      1.95 5.14e- 2  1.04      5.35
```

Metric (continuous) exposure variables I

- The effect of energy on failure, when energy is measured as a continuous variable

R code and output

```
> tidy(glm(chd ~ energy + offset(log(y)), data=diet, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)    0.236  0.976        -1.48 0.139     0.0344      1.58
2 energy         0.999  0.000364     -3.16 0.00159   0.998       1.00
```

- For each 1 unit increase in energy intake, the CHD rate is reduced by 0.1%. The units of energy are kcals per day. The intercept term is the rate when energy intake is zero (which is expected to be outside the observed values).

Metric (continuous) exposure variables II

R code and output

```
> summary(diet$energy)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1748    2537    2803    2829    3110    4396
```

Metric (continuous) exposure variables III

- To get the IRR for an increase of, say, 100 units and centering the effect for energy at 2800 kcals/day:

R code and output

```
> tidy(glm(chd ~ I((energy-2800)/100) + offset(log(y)), data=diet, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term                   estimate std.error statistic   p.value conf.low conf.high
  <chr>                     <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)             0.00940    0.158      -29.6  9.31e-193  0.00677    0.0126
2 I((energy - 2800)/100)  0.891      0.0364      -3.16 1.59e-  3  0.829      0.956
```

- The estimated IRR is $0.99885^{100} = 0.8913$. That is, for each 100 unit increase in energy intake, we estimate that the CHD rate is reduced by 11%. Also the intercept term is the rate when energy is 2800 kcals/day.

The main effects model — constant effect over strata

- If the true effect of exposure does not vary across strata of another variable we can use a main effects model.
- For example, if the estimates of high energy differ only randomly over age, we can consider a model in which the true effect is constant over age, i.e. no interaction.
- This allows us to combine the information from different strata to yield a single estimate of exposure effect.
- This combined estimate of the effect we call *the main effect*, which is then *controlled for* the stratifying (confounding) variable(s).
- Statistical tests for the presence of effect modification are available (although there are no statistical tests for confounding).

Main effects model using Poisson regression

R code and output

```
> tidy(glm(chd ~ hieng + job + offset(log(y)), data=diet, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term          estimate std.error statistic  p.value conf.low conf.high
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)     0.0132     0.312   -13.9   9.41e-44  0.00679    0.0233
2 hienghigh       0.525      0.302    -2.13  3.29e- 2  0.285      0.941
3 jobconductor    1.36       0.393     0.779 4.36e- 1  0.627      2.99
4 jobbank         0.884      0.365    -0.337 7.36e- 1  0.438      1.86
```

- The second row reported is the effect of `hieng` controlled for `job`, and the next two are the effects of `job` controlled for `hieng`.

Models and parameters in Poisson regression I

- In the Poisson regression model we estimated 4 parameters. One parameter (the intercept) is a log rate and the other three are log incidence rate ratios.
- The model is

$$\lambda(hieng, job) = \exp\left(\beta_0 + \beta_1 I(hieng = high) + \beta_2 I(job = cond) + \\ \beta_3 I(job = bank)\right)$$

- $\exp(\beta_0)$ is the predicted rate (not rate ratio) for an individual with *all* covariates at the reference level (i.e. a driver with a low energy intake, or $\lambda(hieng = low, job = driver)$).

Models and parameters in Poisson regression II

The estimated incidence rate ratio for a high energy diet versus a low energy diet is

$$\begin{aligned}\frac{\lambda(hieng = high, job)}{\lambda(hieng = low, job)} &= \frac{\exp\left(\beta_0 + \beta_1 + \beta_2 I(job = cond) + \beta_3 I(job = bank)\right)}{\exp\left(\beta_0 + \beta_2 I(job = cond) + \beta_3 I(job = bank)\right)} \\ &= \exp(\beta_1)\end{aligned}$$

which is independent of the type of job. The estimated incidence rate ratio comparing a conductor to a driver is

$$\begin{aligned}\frac{\lambda(hieng, job = cond)}{\lambda(hieng, job = driver)} &= \frac{\exp\left(\beta_0 + \beta_1 I(hieng = high) + \beta_2\right)}{\exp\left(\beta_0 + \beta_1 I(hieng = high)\right)} \\ &= \exp(\beta_2)\end{aligned}$$

which is independent of high/low energy diet.

Parameter estimates

R code and output

```
> summary(glm(chd ~ hieng + job + offset(log(y)), data=diet, family=poisson))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.3250     0.3118 -13.872   <2e-16 ***
hienghigh     -0.6448     0.3022  -2.134   0.0329 *
jobconductor   0.3063     0.3934   0.779   0.4362
jobbank       -0.1230     0.3651  -0.337   0.7363
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Effect modification

- If the true effect of exposure varies across strata of another variable there is said to be ‘effect modification’ — the effect of exposure cannot then be represented by one IRR.
- Does `job` modify the effect of `hieng`? If we estimate the IRR of high energy separately in all three job groups we get

job	Effect of hieng
driver	0.41
conductor	0.66
bank	0.52

- The figures represent the incidence rate ratios (comparing high to low energy intake) within each job category.
- If the effect of high energy is not modified by job then we would expect these to be similar.

Interaction model using Poisson regression

R code and output

```
> tidy(glm(chd ~ hieng*job + offset(log(y)), data=diet, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term                   estimate std.error statistic  p.value conf.low conf.high
  <chr>                     <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)              0.0145     0.354   -12.0   4.45e-33  0.00661    0.0269
2 hienghigh                0.410      0.612    -1.45  1.46e- 1  0.109      1.30
3 jobconductor             1.14       0.500     0.257 7.98e- 1  0.418      3.09
4 jobbank                  0.813      0.456    -0.452 6.51e- 1  0.337      2.08
5 hienghigh:jobconductor   1.60       0.816     0.573 5.67e- 1  0.325      8.42
6 hienghigh:jobbank        1.26       0.764     0.305 7.61e- 1  0.288      6.06
```

- 0.41 is the effect of `hieng` when `job` is at its first level.
- 1.14 and 0.81 are the effects of `job` when `hieng` is at its first level.
- 1.60 and 1.26 are the interactions between `hieng` and `job`.

Parameters for the interaction model I

We have the regression model

$$\begin{aligned}\lambda(hieng, job) = \exp\big(&\beta_0 + \beta_1 I(hieng = high) + \\ &\beta_2 I(job = cond) + \beta_3 I(job = bank) + \\ &\beta_4 I(job = cond \,\&\, hieng = high) + \\ &\beta_5 I(job = bank \,\&\, hieng = high)\big)\end{aligned}$$

The estimated incidence rate ratio for a high energy diet versus a low energy diet is

$$\begin{aligned}&\frac{\lambda(hieng = high, job)}{\lambda(hieng = low, job)} \\ &= \frac{\exp\left(\beta_0 + \beta_1 + (\beta_2 + \beta_4) I(job = cond) + (\beta_3 + \beta_5) I(job = bank)\right)}{\exp\left(\beta_0 + \beta_2 I(job = cond) + \beta_3 I(job = bank)\right)} \\ &= \exp(\beta_1 + \beta_4 I(job = cond) + \beta_5 I(job = bank))\end{aligned}$$

which is dependent on the type of job (unless $\beta_4 = 0$ and $\beta_5 = 0$).

Parameters for the interaction model II

The estimated incidence rate ratio comparing a conductor to a driver is

$$\begin{aligned}
&\frac{\lambda(hieng, job = cond)}{\lambda(hieng, job = driver)} \\
&= \frac{\exp\left(\beta_0 + (\beta_1 + \beta_4) I(hieng = high)\right)}{\exp\left(\beta_0 + \beta_1 I(hieng = high)\right)} \\
&= \exp(\beta_2 + \beta_4 I(hieng = high))
\end{aligned}$$

which is dependent on high/low energy diet (unless $\beta_4 = 0$).

Testing for interaction using R

- A test of interaction tests if the interactions (β_4 and β_5) are equal to zero.

R code and output

```
> anova(fit)
Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        336     262.82
hieng      1   4.8184       335     258.00  0.02816 *
job        2   1.4781       333     256.52  0.47758
hieng:job  2   0.3332       331     256.19  0.84655
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- No evidence of a statistically significant interaction ($p = 0.85$).
- This is a so-called Wald test, which approximates the likelihood ratio test. We could also use a likelihood ratio test, where we compare the log-likelihoods from the main effects model and the interaction model.

R digression: R formulas

Formula	Equivalent formula
`y ~ a*b`	`y ~ a + b + a:b`
`y ~ (a+b+c)^2 - a:c`	`y ~ a + b + c + a:b + b:c`
`y ~ a - 1`	`y ~ a + 0`

- Moreover, you can use the usual mathematical operators in the formula if you wrap the formula term in `I()` (e.g. `y~I((doe-dob)/365.24)`).
- As a reminder, the term `log(x)` will lead to a parameter being estimated for that term, while the offset term `offset(log(x))` will treat the value as a constant and not estimate a parameter.

Reparameterising the model to directly estimate the effect of exposure in each stratum I

- We are often interested in the effect of the exposure (comparison between high and low energy intake) for each level of the modifier (job).
- We can reparameterise the model to directly estimate parameters of interest (the three IRRs, one for each job).

job	hieng=0	hieng=1
driv	1.0	exp(β_3)
cond	1.0	exp(β_4)
bank	1.0	exp(β_5)

Reparameterising the model to directly estimate the effect of exposure in each stratum II

We have the regression model

$$\begin{aligned}\lambda(hieng, job) = \exp\big(\beta_0 + & \\ & \beta_1 I(job = cond) + \beta_2 I(job = bank) + \\ & \beta_3 I(job = driver \,\&\, hieng = high) + \\ & \beta_4 I(job = cond \,\&\, hieng = high) + \\ & \beta_5 I(job = bank \,\&\, hieng = high)\big)\end{aligned}$$

The estimated incidence rate ratio for a high energy diet versus a low energy diet is

$$\begin{aligned}&\frac{\lambda(hieng = high, job)}{\lambda(hieng = low, job)} \\ &= \frac{\exp\left(\beta_0 + \beta_3 I(job = driver) + (\beta_1 + \beta_4) I(job = cond) + (\beta_2 + \beta_5) I(job = bank)\right)}{\exp\left(\beta_0 + \beta_1 I(job = cond) + \beta_2 I(job = bank)\right)} \\ &= \exp(\beta_3 I(job = driver) + \beta_4 I(job = cond) + \beta_5 I(job = bank))\end{aligned}$$

which is dependent on the type of job (unless $\beta_3 = \beta_4 = \beta_5$).

How to make R produce stratified effects I

- Instead of just one baseline rate, you need three baseline rates, λ (one for each job level). For each job level, you need a IRRs for the energy effect, θ.

R code and output

```
> tidy(glm(chd ~ job + hieng:job + offset(log(y)), data=diet, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term                    estimate std.error statistic  p.value conf.low conf.high
  <chr>                      <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)               0.0145     0.354   -12.0   4.45e-33  0.00661    0.0269
2 jobconductor              1.14       0.500     0.257 7.98e- 1  0.418      3.09
3 jobbank                   0.813      0.456    -0.452 6.51e- 1  0.337      2.08
4 jobdriver:hienghigh       0.410      0.612    -1.45  1.46e- 1  0.109      1.30
5 jobconductor:hienghigh    0.655      0.540    -0.783 4.34e- 1  0.216      1.88
6 jobbank:hienghigh         0.518      0.456    -1.44  1.49e- 1  0.203      1.25
```

- Note that this is the same model; there are still 6 parameters and the fitted values are identical. It's just that the 6 parameters in this model have a different interpretation.

Effects of exposure within each stratum of the modifier

- If we insert the R output in our previous table, we get

job	hieng=0	hieng=1
driv	1.0	0.41
cond	1.0	0.66
bank	1.0	0.52

- The stratum-specific IRRs are similar, there is no evidence of interaction.

Linear combinations of parameters I

- As an alternative to reparameterising the interaction model we can use the new biostat3::lincom() function to estimate the effect of exposure within each level of the modifier together with confidence intervals. Here again is the interaction model with the default parameterisation.

R code and output

```
> tidy(fit <- glm(chd ~ hieng*job + offset(log(y)), data=diet, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term                     estimate std.error statistic  p.value conf.low conf.high
  <chr>                       <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)                0.0145     0.354   -12.0   4.45e-33  0.00661    0.0269
2 hienghigh                  0.410      0.612    -1.45  1.46e- 1  0.109      1.30
3 jobconductor               1.14       0.500     0.257 7.98e- 1  0.418      3.09
4 jobbank                    0.813      0.456    -0.452 6.51e- 1  0.337      2.08
5 hienghigh:jobconductor     1.60       0.816     0.573 5.67e- 1  0.325      8.42
6 hienghigh:jobbank          1.26       0.764     0.305 7.61e- 1  0.288      6.06
```

Linear combinations of parameters II

- The effect of `hieng` for drivers is 0.41. We now estimate the effect of `hieng` for the other two categories of job.

R code and output

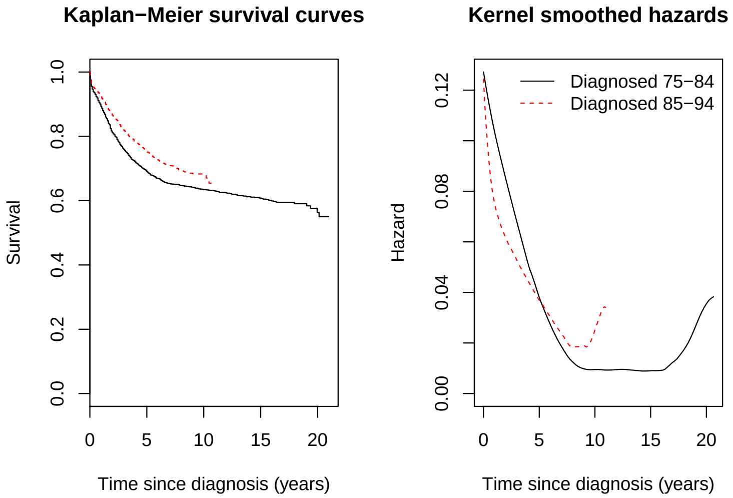
```
> lincom(fit, c("hienghigh+hienghigh:jobconductor",
                "hienghigh+hienghigh:jobbank"),
         eform=TRUE)

                                 Estimate  2.5 %     97.5 %   Chisq    Pr(>Chisq)
hienghigh+hienghigh:jobconductor 0.6550924 0.2273009 1.888008 0.61341  0.4335067
hienghigh+hienghigh:jobbank      0.5177431 0.211639  1.266581 2.079971 0.149243
```

- The calculation $0.410 \times 1.596 = 0.655$ isn't difficult but calculating the standard error and CI is non-trivial (a combination of variances and covariances).

Time varying rates

- So far, we have modelled the overall rate, i.e. a constant rate throughout the follow-up.
- We have modelled how this overall rate could vary according to other variables using main effects models and interaction models.
- Now, we will look at how to model and adjust for time when it confounds the effect of interest.
- The elegant way we can model time is one of the beauties of survival analysis.
- We will look at time as a confounder of the rates and time as an effect-modifier of other variables (later on).

For which calendar period is mortality lowest? I



For which calendar period is mortality lowest? II

R code and output

```
> survRate(Surv((exit - dx)/365.24/1000, status=="Dead: cancer") ~ year8594,
           data=colon,
           subset=stage=="Localised")
                            tstop event     rate    lower    upper
year8594=Diagnosed 75-84 20.33538   905 44.50372 41.65109 47.50027
year8594=Diagnosed 85-94 15.27233   829 54.28118 50.64854 58.10556
```

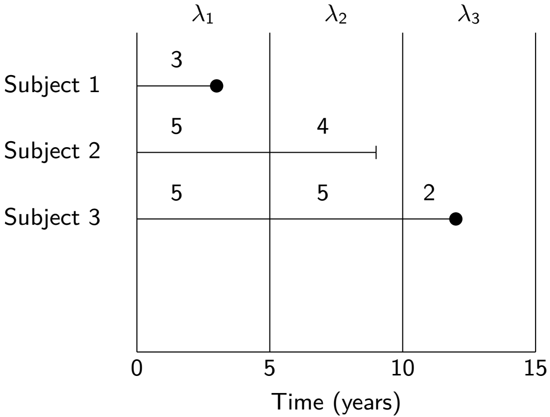
- The graphs suggest that patients diagnosed in the recent period have lower mortality (better survival) but the estimated rates suggest otherwise.
- The end of follow-up is 1995. Those diagnosed 1975-84 are followed for up to 20 years, whereas those diagnosed 1985-94 are followed for at most 10 years.
- Those diagnosed 1985-94 have shorter follow-up.
- Since the effect is highest in the early years of follow-up, the rates are confounded by follow-up time.

Time as a confounder I

- When the rate changes with time then time may confound the effect of exposure.
- We will, for the moment, assume that the rates are constant within broad time bands but can change from band to band.
- This approach (categorising a metric variable and assuming the effect is constant within each category) is standard in epidemiology.
- We often categorise metric variables — the only difference here is that the variable is 'time'.

Time as a confounder II

- Consider a group of subjects with rates λ_1 during band 1 (0-5 years), λ_2 during band 2 (5-10 years), etc.



- What are the estimated failure rates, λ_1, λ_2, λ_3, for each of the bands?

Splitting the records by follow-up time

- A convenient way to fit these models using a computer is to replace the single record for this subject by three new records, one for each band of observation.
- The new subject–band records can be treated as independent records.

subject	timeband	follow-up	failure
1	0-5	3	1
2	0-5	5	0
2	5-10	4	0
3	0-5	5	0
3	5-10	5	0
3	10-15	2	1

- The rate for timeband 0-5 is then $1/(3+5+5)$, and so on for other time bands.
- This method can be used whether rates are varying simply as a function of time or in response to some time–varying exposure.

Splitting on ‘time in study’ (time since entry) I

- It is good to check the data before splitting!

R code and output

```
> diet <- transform(diet, surv=Surv((dox-doe)/365.24, chd))
> subset(diet, id==34)

    id chd        y hieng  energy    job month height  weight        doe
138 34   1 7.709788   low 2561.83 driver     4  177.8 66.4524 1959-04-16
           dox        dob      yoe      yox      yob     surv
138 1966-12-31 1899-06-11 1959.287 1966.997 1899.441 7.709999
```

Splitting on ‘time in study’ (time since entry) II

- Split the data using the `survival::survSplit` function:

R code and output

```
> diet2 <- survSplit(Surv((dox-doe)/365.24, chd) ~ ., diet,
                     cut=seq(2,20,by=2),
                     episode="timeband")
> subset(diet2, id==34, select=c(id, tstart, tstop, chd, timeband))

    id tstart    tstop chd timeband
966 34      0 2.000000   0        1
967 34      2 4.000000   0        2
968 34      4 6.000000   0        3
969 34      6 7.709999   1        4
```

- Person ID=34 was followed up for 7.71 years, and when we split the record we got four rows of data, one for each time band 0–2, 2–4, 4–6 and 6–8 years where this person contributes risk time.

Rates for different time bands I

R code and output

```
> survRate(Surv(tstart,tstop,chd)~timeband, data=diet2)
            timeband      tstop event        rate        lower       upper
timeband=1         1 665.7757091     6 0.009012044 3.307261e-03  0.019615426
timeband=2         2 649.9296901     3 0.004615884 9.519062e-04  0.013489572
timeband=3         3 618.7236885    11 0.017778534 8.874980e-03  0.031810708
timeband=4         4 594.7178841     8 0.013451756 5.807514e-03  0.026505322
timeband=5         5 566.9969335     1 0.001763678 4.465246e-05  0.009826585
timeband=6         6 491.9313328     8 0.016262432 7.020964e-03  0.032043475
timeband=7         7 414.8095499     2 0.004821490 5.839048e-04  0.017416879
timeband=8         8 361.9300186     5 0.013814825 4.485636e-03  0.032239194
timeband=9         9 177.8258679     2 0.011246958 1.362059e-03  0.040627878
timeband=10       10  60.9876246     0 0.000000000 0.000000e+00  0.060485705
timeband=11       11   0.1664659     0 0.000000000 0.000000e+00 22.159972563
```

Rates for different time bands II

- Poisson regression can also be performed using the `glm` function.

R code and output

```
> summary(fit <- glm(chd~hieng+offset(log(tstop - tstart)),
                     data=diet2, family=poisson))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2980     0.1890 -22.743   <2e-16 ***
hienghigh    -0.6532     0.3021  -2.162   0.0306 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> tidy(fit, conf.int=TRUE, exponentiate=TRUE)
  term        estimate std.error statistic   p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)   0.0136     0.189    -22.7  1.69e-114  0.00916    0.0193
2 hienghigh     0.520      0.302     -2.16 3.06e-  2  0.283      0.933
```

Rates for different time bands III

- The effect of `hieng` controlled for `timeband` is found with:

Rates for different time bands IV

R code and output

```
> fit <- glm(chd~hieng+offset(log(tstop - tstart))+
                 factor(timeband),
             data=diet2, family=poisson)
> ## tidy(fit, conf.int=TRUE, exponentiate=TRUE) # fails:(
> biostat3::eform(fit)

                       exp(beta)       2.5 %     97.5 %
(Intercept)         1.220474e-02 0.005313288 0.02803458
hienghigh           5.192323e-01 0.287195398 0.93874145
factor(timeband)2   5.135333e-01 0.128433239 2.05333504
factor(timeband)3   1.994008e+00 0.737406016 5.39196399
factor(timeband)4   1.509798e+00 0.523838090 4.35151497
factor(timeband)5   1.977773e-01 0.023810314 1.64281217
factor(timeband)6   1.808344e+00 0.627449679 5.21174345
factor(timeband)7   5.338991e-01 0.107759513 2.64522605
factor(timeband)8   1.535991e+00 0.468770132 5.03288826
factor(timeband)9   1.261453e+00 0.254598705 6.25008866
factor(timeband)10 1.102078e-06 0.000000000        Inf
factor(timeband)11 3.508572e-05 0.000000000        Inf
```

Rates for different time bands V

- We could also use either timeband or the interval mid-point as a linear effect:

R code and output

```
> diet3 <- transform(diet2,tmid=(tstop+tstart)/2)
> broom::tidy(glm(chd~hieng+offset(log(tstop - tstart))+tmid,
              data=diet3, family=poisson),
       conf.int=TRUE, exponentiate=TRUE)
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)   0.0155    0.293    -14.2   4.78e-46 0.00845     0.0267
2 hienghigh     0.521     0.302     -2.16  3.07e- 2 0.283       0.933
3 tmid          0.983     0.0308    -0.560 5.75e- 1 0.924       1.04
```

- Moreover, we could also model for the time mid-points using splines or some other smooth function (we will discuss splines further on day 4).
- There is no reason to believe that time-on-study would be a confounder for these data. This would, however, be of interest in the cancer examples.

Splitting the follow–up on the age scale I

- Attained age is a possible confounder for the diet study, young and old people may differ in both energy intake and risk for CHD. Attained age is more interesting as a potential confounder than age at entry.

R code and output

```
> diet2 <- survSplit(Surv((doe-dob)/365.24, (dox-dob)/365.24, chd) ~ .,
                     data=diet,
                     cut=seq(40,70,by=10),
                     episode="ageband")
> subset(diet2, id==34, select=c(id, tstart, tstop, chd, ageband))

    id   tstart    tstop chd ageband
333 34 59.84558 60.00000   0       3
334 34 60.00000 67.55558   1       4
```

Splitting the follow–up on the age scale II

- We see that, as expected, the CHD incidence rate depends on attained age.

R code and output

```
> survRate(Surv(tstart,tstop,chd)~ageband, data=diet2)
          ageband        tstop event        rate       lower       upper
ageband=1       1 9.630271e+01     0 0.000000000 0.000000000  0.03830505
ageband=2       2 9.068131e+02     6 0.006616579 0.002428168  0.01440151
ageband=3       3 2.106989e+03    18 0.008542995 0.005063120  0.01350162
ageband=4       4 1.493650e+03    22 0.014729023 0.009230600  0.02229992
ageband=5       5 3.997372e-02     0 0.000000000 0.000000000 92.28262547
```

The effect of hieng controlled for attained age I

R code and output

```
> diet2 <- survSplit(Surv((doe-dob)/365.24, (dox-dob)/365.24, chd) ~ .,
                     data=diet,
                     cut=seq(50,60,by=10),
                     episode="ageband")
> fit <- glm(chd~hieng+offset(log(tstop - tstart))+
                 factor(ageband),
             data=diet2, family=poisson)
> tidy(fit, conf.int=TRUE, exponentiate=TRUE)
  term             estimate std.error statistic  p.value conf.low conf.high
  <chr>               <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)       0.00840     0.432   -11.1   1.86e-28  0.00321    0.0179
2 hienghigh         0.536       0.303    -2.06  3.94e- 2  0.291      0.962
3 factor(ageband)2  1.35        0.472     0.641 5.22e- 1  0.567      3.74
4 factor(ageband)3  2.33        0.461     1.83  6.70e- 2  1.00       6.33
```

- Is there evidence that the effect of hieng is confounded by attained age?

Collapsing person-time data I

- Fine time-splitting on one or more time scales can make the dataset very large.
- Usefully, for time segments with the same covariates, we can also collapse, where we sum the events and sum the person-time

R code and output

```
> diet3 <- group_by(diet2, ageband, hieng) |>
      summarise(chd=sum(chd), pt=sum(tstop-tstart)) # n=6 observations
> fit <- glm(chd~hieng+offset(log(pt))+
             factor(ageband),
             data=diet3, family=poisson)
> tidy(fit, conf.int=TRUE, exponentiate=TRUE) # essentially the same fit
  term             estimate std.error statistic  p.value conf.low conf.high
  <chr>               <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)       0.00840     0.432   -11.1   1.88e-28  0.00321    0.0179
2 hienghigh         0.536       0.303    -2.06  3.95e- 2  0.291      0.962
3 factor(ageband)2  1.35        0.472     0.641 5.22e- 1  0.567      3.74
4 factor(ageband)3  2.33        0.461     1.83  6.70e- 2  1.00       6.33
```

Splitting by multiple time scales I

- Usefully, we can also split by multiple time scales.

R code

```
## first, split by attained age
diet2 <- survSplit(Surv((doe-dob)/365.24, (dox-dob)/365.24, chd) ~ .,
                   data=transform(diet,yob=year(dob)),
                   cut=seq(50,60,by=10), # 40-, 50-, 60-
                   episode="ageband")
## then split by attained year
diet3 <- survSplit(Surv(yob+tstart, yob+tstop, chd) ~ .,
                   data=diet2,
                   cut=seq(1960,1980,by=10), # 1950-, 1960, 1970-
                   episode="yearband")
## then fit the regression model
fit <- glm(chd~hieng+offset(log(tstop - tstart))+
               factor(ageband)+factor(yearband),
           data=diet3, family=poisson)
summary(fit)
```

Splitting by multiple time scales II

R output

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.9046     0.5628  -6.937    4e-12 ***
hienghigh          -0.6030     0.3033  -1.988   0.0468 *
factor(ageband)2    0.3958     0.4775   0.829   0.4072
factor(ageband)3    1.0530     0.5003   2.105   0.0353 *
factor(yearband)2  -1.0708     0.5029  -2.129   0.0332 *
factor(yearband)3  -1.1298     0.5509  -2.051   0.0403 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Statistical models I

- Multiple regression models are important in that they allow simultaneous estimation and testing of the effect of many prognostic factors on survival.
- The aim of statistical modelling is to derive a mathematical representation of the relationship between an observed response variable and a number of explanatory variables, together with a measure of the uncertainty of any such relationship.
- The uses of a statistical model can be classified into the following three areas:
  1. Descriptive: To describe any structure in the data and quantify the effect of explanatory variables, and to study the pattern of any such associations;
  2. Hypothesis testing: To statistically test whether an observed response variable is associated with one or more explanatory variables; and
  3. Prediction: For example, predicting excess mortality for a future time period, or predicting the way in which the outcome may change if certain explanatory variables changed in value.
- Note that a statistical model is never true, but may be useful.
- When making inference based on the model we assume that the model is true.

Statistical models II

- If the model is badly misspecified then inference will be erroneous.
- It is therefore important to consider the validity of any assumptions (e.g. proportional hazards) underlying the model and to check for evidence of lack-of-fit.

Summary of Day 2

- A *rate* is defined as *events* divided by *total time-at-risk*, where time at risk is usually measured in person-years, person-months etc.
- The rates can vary across various *time scales*, e.g. time since entry, attained age, calendar period.
- Rates can be modelled using Poisson regression, which estimates the baseline hazard rate and the incidence rate ratios (IRR) for different exposure levels.
- Interactions can be re-parameterised in various ways to show the different aspects of effect modification.
- The estimates of rates and IRRs can be confounded by time.

Exercises for Wednesday afternoon

6. Diet data: tabulating incidence rates and modelling with Poisson regression. [working through what was presented in the lectures].
7. Localised melanoma: model cause-specific mortality with Poisson regression. [this is a key exercise, tomorrow we will fit a Cox model to the same data and compare the results]
8. Diet data: Using Poisson regression to study the effect of energy intake adjusting for confounders. [Something for you to do if you've finished the other two]

References I

John P. Klein and Melvin L. Moeschberger.
Survival Analysis: Techniques for Censored and Truncated Data.
Springer-Verlag, 1997.