Solution to the practicals - matching

Epidemiological methods in medical research 2023

9 March 2023

Exercise 1: Warming-up

1. **i.d** retrospective case control study is (usually) well suited to study rare outcomes.

   i.f retrospective case control study may be subject to survivor bias.

   ii.b matching partially controls for known and unknown confounding.

   ii.g matching may induce over-adjustment (adjustment for a variable that is a mediator).

   iii.a randomization controls for known and unknown confounding.

   iii.e randomization is not ethical to use when the exposure is known to be harmful.

   iv.c stratification controls for known confounding.

   iv.g stratification may induce over-adjustment (adjustment for a variable that is a mediator)

2. This would be a matched case–control study.

   - by using the other twin as a matched control we would typically adjust for age, genetic factors (perfectly for monozygotic twins and partially for dizygotic twins), calendar year, sex (if same sex twins), some familial and environmental factor[1].
   - there could be confounding due to environmental factors, e.g. exposure to chemicals if the twins do not have the same job, access to the health care system if they leave in different regions. Prevention, e.g. physical exercise may also differ between twins.
   - Studying the effect of the diet on obesity in pre-adult twins would not be a very efficient design. Indeed twins are likely to follow the same diet (or a similar diet) as they often share meals and are used to the same type of food. Only twins with different diets would bring substantial information.

[1] note that is not easy to precisely list the variables we adjust on

Exercise 2: Food poisoning in Fyn

1. a) The odd ratio can be computed using the formula seen in lecture 3: $OR = (ad)/(bc)$

```
a <- t2x2["0","control"]
b <- t2x2["0","case"]
c <- t2x2["1","control"]
d <- t2x2["1","case"]
(a*d) / (b*c)
```

```
[1] 4.861111
```

1. b) No because the cases and controled were matched and we need to account for that in the analysis.

1. c) We can see that in each set there either 1 or 0 case and 2 or 3 controls. We don't expect the set 10 and 22 to be informative since they do not contain any cases (i.e. they could be ignored in the analysis).

2. This logistic regression gives the same odd ratio as our "manual" calculation. It is also incorrect as it does not account for the matching.

3. a) We now account for the matching using a conditional logistic regression: we only compare cases and controls belonging to the same set. The strata statement indicates how the cases and control were matched.

3. b) The estimated odd ratio is 4.61 [1.19;17.79] witha p-value of 0.026 so there is some evidence for a harmful effect of being exposed to `hamburg`. The magnitude of the effect is unclear as the confidence interval is very wide.

3. c) As discussed previously, there is no case in the set `10` so removing this set will not affect the estimates:

```
m1.bis <- clogit( caco=="case" ~ factor(hamburg) + strata(set),
                  data=manh[manh$set!=10,] )
ci.exp( m1.bis, pval = TRUE )
```

```
                  exp(Est.)     2.5%    97.5%          P
factor(hamburg)1  4.617469 1.198331 17.79226 0.02622569
```

4. a) We first initialize a list to store the conditional logistic models relative to each exposure (`ls.logit`), and a matrix to store the estimate odd ratios, confidence intervals and p-values (`M.cOR`).
   We then loop over each exposure. First we create the appropriate formula (`iFormula`), then display the 2 by 2 table corresponding to the exposure (`table(...)`), fit the conditional logistic model (`clogit(...)`) and save the odd ratio, its confidence intervals and p-value in the right line of `M.cOR`.
   The `cat` function enable to print what happens as R execute the loop.

5. b) For two exposures (`lamkod` and `rgtmbr`) where the was no exposed individuals: it was impossible for the conditional logistic model to estimate an odd ratio. This is why the results is NA, i.e., we don't know as we have no information from the data.
   We got warnings from two exposures where there was no case that were not exposed (`svinkod` and `kodpaal`). It is thus not possible to estimate the effect and so the software returned warnings and very wide confidence intervals. The estimated odd ratio should therefore not be taken seriously.
   We also got warnings from the exposure "filet". The 2 by 2 table seems "normal", which is confirmed by the convergence of the logistic model:

```
glm(caco.bin ~ filet, family = binomial, data = manh)
```

```
Call:  glm(formula = caco.bin ~ filet, family = binomial, data = manh)

Coefficients:
(Intercept)        filet
     -1.969        2.886

Degrees of Freedom: 62 Total (i.e. Null);  61 Residual
Null Deviance:            71.4
Residual Deviance: 53.19         AIC: 57.19
```

So the issue must come from conditionning on `set`. We can explicit the 2 by 2 tables stratified on `set`:

```
table3D <- table(manh$caco,manh$filet,manh$set)
ftable(table3D)
```

```
           1 3 4 5 7 8 9 10 11 12 14 16 17 18 19 20 22 23

control 0  2 1 1 3 3 2 3  2  3  2  3  3  3  3  3  3  0  3
        1  0 1 0 0 0 0 0  0  0  1  0  0  0  0  0  0  2  0
case    0  1 0 0 0 0 1 0  0  0  0  1  0  1  1  0  1  0  0
        1  0 1 1 1 1 0 1  0  1  1  0  1  0  0  1  0  0  1
```

we see that there is no set with exposed control and non-exposed case. The common odd ratio via the Mantel-Haenszel method would be infinite:

```
mantelhaen.test(table3D.dis)
```

```
        Mantel-Haenszel chi-squared test with continuity correction

data:  table3D.dis
Mantel-Haenszel X-squared = 18.188, df = 1, p-value = 2.002e-05
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 NaN NaN
sample estimates:
common odds ratio
              Inf
```

which is in line with the fact that the conditional logistic model does not converge.

Note: this illustrates that when asking something "impossible" to do, the software may react differently (depending on how it has been program): return an error, a warning, or weird estimates. So one need to be careful and be critical when seeing very large confidence intervals and standard errors (the p-value does not help us here).

c) This approach is exploratory in the sense that we test many possiblities without much a priori. If we were to focus on the exposure that is the "most significant" or has the "largest effect", we would need to account for multiple testing. In that case, the effect observed in `hamburg` is not be very convincing and we would ask for more evidence for concluding about `hamburg`. This is quite different from what we concluded in the previous question when we a priori focused on `hamburg`.

Exercise 3: IHD data from Clayton & Hills

1. We fit the Poisson model as described in the question:

```
m1 <- glm(cbind(chd,y) ~ age + energy, family = poisreg, data = diet )
ci.exp( m1, pval = TRUE )
```

```
             exp(Est.)         2.5%     97.5%           P
(Intercept) 0.01290045 0.0005247771 0.3171280 0.007746408
age         1.05428448 1.0067255164 1.1040902 0.024795165
energy      0.90065355 0.8381702851 0.9677948 0.004340137
```

and see a clear effect of energy intake on mortality - the higher the energy intake the lower the mortality, one unit difference is associated with a RR of 0.9. To have an idea of how "big" is one unit difference in energy intake, we either use prior knowledge or have a look to the empirical distribution of energy intake:

```
quantile( diet$energy, probs = c(0,0.25,0.5,0.75,1) )
```

```
     0%     25%     50%     75%    100%
17.4843 25.3669 28.0298 31.0966 43.9575
```

2. We update the model using linear splines:

```
m2 <- glm(cbind(chd,y) ~ age + energy + pmax(age-45,0) + pmax(age-53,0) +
    pmax(age-61,0),
           family = poisreg, data = diet )
ci.exp( m2, pval = TRUE)
```

```
                    exp(Est.)         2.5%      97.5%           P
(Intercept)       0.003690826 8.848424e-07 15.3950555 0.187795038
age               1.089323162 9.045658e-01  1.3118171 0.366928328
energy            0.899899607 8.370428e-01  0.9674765 0.004304294
pmax(age - 45, 0) 0.914143918 6.944319e-01  1.2033709 0.522152555
pmax(age - 53, 0) 1.100809278 8.462433e-01  1.4319536 0.474127015
pmax(age - 61, 0) 1.159395384 7.106399e-01  1.8915314 0.553715081
```

Here we have used `pmax` which is the function that takes the maximum of two vectors component-wise (parallel maximum). We see that the effect of energy does not change much if age is modeled as a linear spline:

```
rbind(linear = ci.exp( m1 )["energy",],
      spline = ci.exp( m2 )["energy",])
```

```
       exp(Est.)      2.5%     97.5%
linear 0.9006536 0.8381703 0.9677948
spline 0.8998996 0.8370428 0.9674765
```

3. We now use linear splines for energy instead of age:

```
m3 <- glm(cbind(chd,y) ~ age + energy + pmax(energy-20,0) + pmax(energy
   -30,0),
          family = poisreg, data = diet )
ci.exp( m3, pval = TRUE )
```

```
                        exp(Est.)         2.5%        97.5%          P
(Intercept)          0.0009209113 1.216440e-12 6.971798e+05 0.50278537
age                  1.0543291791 1.006751e+00 1.104155e+00 0.02473162
energy               1.0275238222 3.685926e-01 2.864423e+00 0.95860217
pmax(energy - 20, 0) 0.8800115046 3.011375e-01 2.571650e+00 0.81528250
pmax(energy - 30, 0) 0.9651885719 7.060877e-01 1.319367e+00 0.82418596
```

There is no evidence of any non-linear effect of energy:

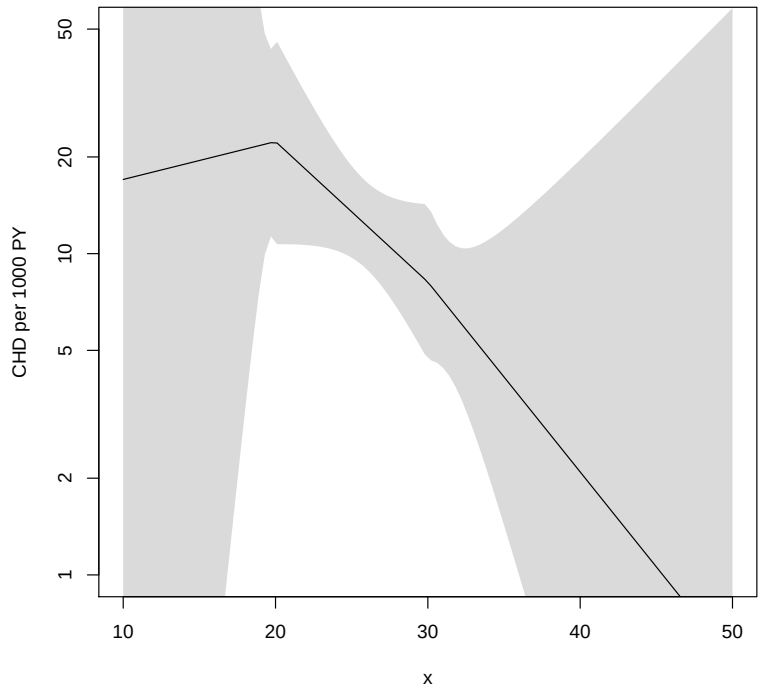
```
anova( m1, m3, test="Chisq" )
```

```
Analysis of Deviance Table

Model 1: cbind(chd, y) ~ age + energy
Model 2: cbind(chd, y) ~ age + energy + pmax(energy - 20, 0) + pmax(energy -
    30, 0)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       334     247.22
2       332     247.06  2  0.15662   0.9247
```

Note 1: we can use the `ci.pred` function to display the model fit for 50 year old men at different levels of energy:

```
nd <- data.frame(age = 50, energy = seq(10,50,length=100))
pr <- ci.pred(m3, nd)
matshade(x = nd$energy, y = pr*1000, log="y", plot=TRUE, ylim=c(1,50),
   ylab="CHD per 1000 PY")
```



Note 2: in the previous approach we combined linear functions to model non-linearities and had to decide on the position of the knots. It is possible to use a more flexible and data-driven approach based on thin plate regression. Here we don't have to specify knots:

```
library(mgcv)
m3.bis <- gam(chd ~ age + energy + s(energy, m=c(2, 0)), method = "REML",
              family = poisson, data = diet, offset = log(y))
summary(m3.bis)
```

```
Indlæser krævet pakke: nlme
This is mgcv 1.8-40. For overview type 'help("mgcv-package")'.

Family: poisson
Link function: log

Formula:
chd ~ age + energy + s(energy, m = c(2, 0))

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.35046    1.63388  -2.663  0.00775 **
age          0.05286    0.02355   2.245  0.02480 *
energy      -0.10464    0.03669  -2.852  0.00435 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df Chi.sq p-value
s(energy) 0.0001064      8      0   0.919

R-sq.(adj) =  -0.165   Deviance explained = 5.94%
-REML = 175.83  Scale est. = 1         n = 337
```

It also outputs a test for the non-linearity of the enery effect that is not significant (p=1).

4. We fit the Cox model using:

```
mc <- coxph( Surv(y, chd) ~ energy, data = diet )
```

and see remarkably little difference between the two approaches:

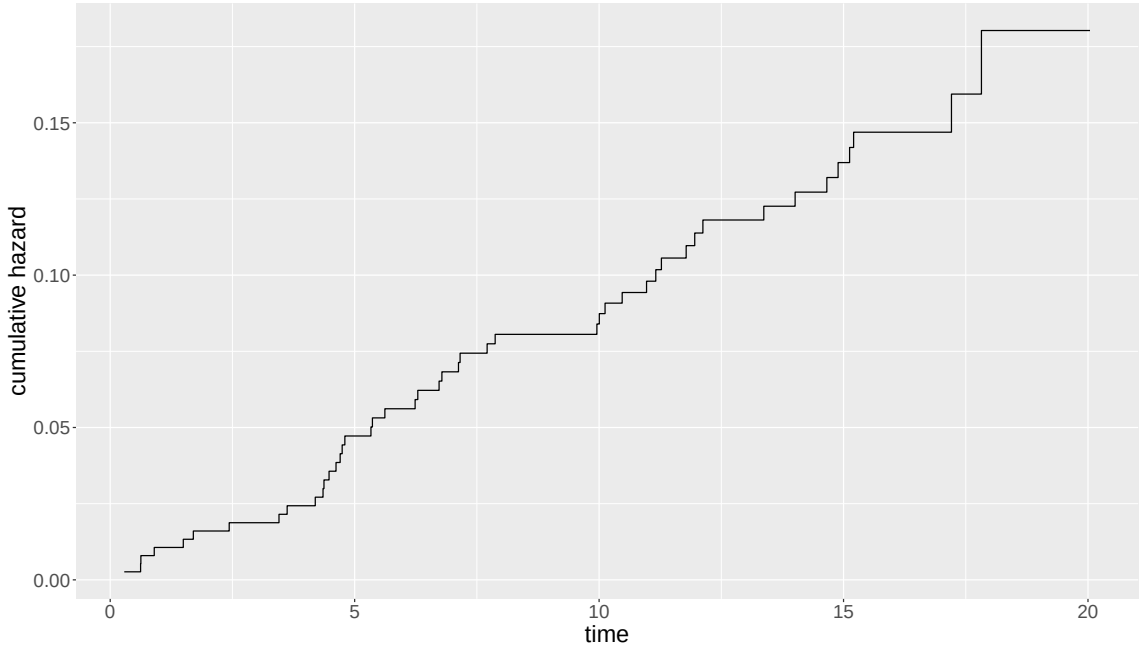
```
rbind( ci.exp( mc, pval = TRUE ),
       ci.exp( m2, subset="energy", pval = TRUE) )
```

```
       exp(Est.)      2.5%     97.5%           P
energy 0.8922424 0.8307192 0.9583221 0.001761108
energy 0.8998996 0.8370428 0.9674765 0.004304294
```

This is line with the fact that the hazard appears to be relatively constant over time as the cumulative hazard increases almost linearly over time:

```
mc.lambda <- basehaz(mc)
library(ggplot2)
ggplot(mc.lambda, aes(x=time, y=hazard)) + geom_step() + labs(y = "
    cumulative hazard")
```

```
Advarsel: pakke 'ggplot2' blev bygget under R version 4.2.2
```



5. In R we can compute age at exit as `age+py` on the fly and use in the Cox model directly:

```
ma <- coxph( Surv(age, age+y, chd) ~ energy, data = diet )
rbind(cox_py = ci.exp( mc, pval = TRUE ),
      cox_age = ci.exp( ma, pval = TRUE ),
      poisson = ci.exp( m1, subset="energy", pval = TRUE) )
```

```
       exp(Est.)      2.5%     97.5%           P
energy 0.8922424 0.8307192 0.9583221 0.001761108
energy 0.8965089 0.8342417 0.9634237 0.002934558
energy 0.9006536 0.8381703 0.9677948 0.004340137
```

The introduction of current age as the underlying time scale does not change much the estimates.