Solution to the Practicals - Case control study

Epidemiological methods in medical research 2023

9 February 2023

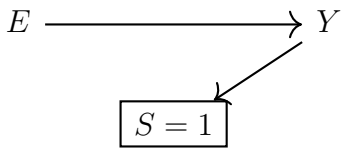
Exercise 1: Warm-up about case-control design

1.1 Specificities of a case control study

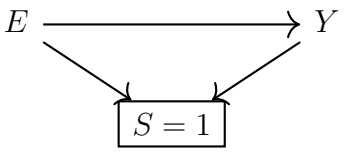
1. In a cohort study, the relationship between exposure and disease incidence is investigated by **following the entire cohort** and measuring the rate of occurrence of new cases in the different exposure groups. In contrast, in a case-control study we do not consider the entire cohort. We typically consider all the cases (sick individuals) and **sample a subset of controls** (disease-free individual).

2. By sampling less controls we will increase the disease frequency (prevalence or risk) which means that the prevalence of the disease (or the risk of a disease) cannot be estimated without bias. The same applies for the risk difference and risk ratio. Only the odds ratio can be estimated without bias.

3. Correct sampling of the controls is key to obtain unbiased estimates. In particular the selection probabilities for controls should not vary between exposure groups. Selection bias will occur when this is not true. This can be illustrated using a simple DAG where $S = 1$ denote being included in the study:



DAG of a case control study where sampling probabilities do not depend on the exposure meaning that we can estimate the association between E and Y. It does not mean that all estimation methods will be valid though, as only the odds ratio will provide reasonable estimates.



DAG of a case control study where sampling probabilities depend on the exposure. It induces a collider bias meaning that we cannot estimate (correctly) the association between E and Y.

A typical threat to this assumption is differential compliance, e.g. if many controls refuse to participate especially when they have not been subject to the exposure of interest.
Difficulties will also arise when using a hospital-based case control study:

- cases are new patients diagnosed at the hospital with the disease of interest
- controls are hospitalized patients with other diseases

However patients sick with other diseases are not representative of persons free of the disease of interest - they will typically have specific risk factors (which will appear protective for the disease of interest as they are over-represented among the controls).
Importantly controls represent risk time not persons immune to the disease. Typically they are only disease free up to when cases experience the disease but may very well experience the disease later on.

1.2 Case control as a cohort study

1. a) The probility of being exposed and die equals the probability of being exposed ($p = 0.1$) times the probability of dying when having been exposed ($\pi_1 = 0.2$) so we would expect $1000 * p * \pi_1 = 20$ persons.
   Similarly we would expect $1000 * p * (1 - \pi_1) = 80$ persons to stay alive while having been exposed, $1000 * (1 - p) * \pi_0 = 45$ persons to die while not having been exposed, and $1000 * (1 - p) * (1 - \pi_0) = 855$ persons to stay alive while not having been exposed

1. b) The risk of death among the exposed would be the number of death among the exposed divided by the number of exposed: $r_1 = \frac{1000p\pi_1}{1000p} = \pi_1 = 0.2$. Similarly the risk among the non-exposed is $r_0 = \frac{1000(1-p)\pi_0}{1000(1-p)} = \pi_0 = 0.05$. This leads to a risk difference of 0.15, risk ratio of 4, and odds ratio of:

$$OR = \frac{r_1/(1-r_1)}{r_0/(1-r_0)} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{0.2/0.8}{0.05/0.95} = 4.75 \tag{1}$$

2. a) The probility of being included, exposed, and die equals the probability of being exposed ($p = 0.1$) times the probability of dying when having been exposed ($\pi_1 = 0.2$) times the probability of being included when exposed and dead $s_{1,\text{case}}$ so we would expect $1000 * p * \pi_1 = 20$ persons to die while having been exposed and included.
   Similarly we would expect $1000 * p * (1 - \pi_1) * s_{1,\text{control}} = 8$ persons to stay alive while having been exposed and included, $1000 * (1 - p) * \pi_0 * s_{0,\text{case}} = 45$ persons to die while not having been exposed and included, and $1000 * (1 - p) * (1 -$

$\pi_0) * s_{0,\text{control}} = 85.5$ persons to stay alive while not having been exposed and included.

2. b) The risk of death among the exposed would be the number of death among the exposed divided by the number of exposed:

$$r_1' = \frac{1000p\pi_1 s_{1,\text{case}}}{1000p\pi_1 s_{1,\text{case}} + 1000p(1-\pi_1)s_{1,\text{control}}} = \frac{20}{20+8} \approx 0.7143$$
$$r_1' = \frac{\pi_1}{\pi_1 + (1-\pi_1)s_{1,\text{control}}/s_{1,\text{case}}} \tag{2}$$

Similarly the risk among the non-exposed is

$$r_0' = \frac{1000(1-p)\pi_0 s_{0,\text{case}}}{1000(1-p)\pi_0 s_{0,\text{case}} + 1000(1-p)(1-\pi_0)s_{0,\text{control}}} = \frac{45}{45+85.5} \approx 0.3448$$
$$r_0' = \frac{\pi_0}{\pi_0 + (1-\pi_0)s_{0,\text{control}}/s_{0,\text{case}}} \tag{3}$$

This leads to a risk difference of 0.369, risk ratio of 2.071, and odds ratio of

$$OR' = \frac{r_1'/(1-r_1')}{r_0'/(1-r_0')} = \frac{\pi_1/((1-\pi_1)s_{1,\text{control}}/s_{1,\text{case}})}{\pi_0/((1-\pi_0)s_{0,\text{control}}/s_{0,\text{case}})} = \frac{20/8}{45/85.5} = 4.75$$
$$= OR\frac{s_{1,\text{case}}/s_{1,\text{control}}}{s_{0,\text{case}}/s_{0,\text{control}}} \tag{4}$$

2. c) When keeping all cases and subsampling the controls, we bias the risk toward higher risks which will in turn bias the risk difference and the risk ratio. However the odd ratio is unaffected as soon as the sampling probabilities for controls are the same among the exposed and non-exposed, i.e. $s_{1,\text{control}} = s_{0,\text{control}}$.
More precisely equations (2) and (3) show that the risk can only be estimated correctly when the sampling probability for the cases equals that of the control in a given exposure group.
Equation (4) show that unbiased estimation of the odds ratio requires that the ratio between the sampling probabilities cases vs. controls is the same between exposure groups. This is true when the sampling probabilities of the cases and the controls are independent of the prevalence (sufficient but not necessary condition).
It would be a bad idea: since the sampling probability for the control would dependent on the exposure, we would not be able to get an unbiased estimate of the odd ratio. Note that because the sampling probabilities are identical among the exposed, we would be able to estimate the risk among the exposed but we would have nothing to compare it with.

Exercise 2: Case study: BCG study

1. a) We fit a logistic model with an interaction age and vaccination status:

```
summary(e.glmAllI)$coef
```

```
               Estimate Std. Error    z value     Pr(>|z|)
(Intercept)  -8.9349820  1.0000658 -8.9343937 4.094169e-19
age2          2.4589892  1.0445959  2.3540099 1.857212e-02
age3          3.6347023  1.0178527  3.5709511 3.556874e-04
age4          4.0077284  1.0310599  3.8869984 1.014914e-04
age5          4.1317810  1.0249594  4.0311653 5.550098e-05
age6          4.1391915  1.0139719  4.0821560 4.461984e-05
age7          4.2200991  1.0107418  4.1752495 2.976596e-05
age1:scarYes -0.4339847  1.4142903 -0.3068569 7.589523e-01
age2:scarYes -0.1135229  0.4032064 -0.2815505 7.782882e-01
age3:scarYes -0.5394366  0.2854458 -1.8898039 5.878419e-02
age4:scarYes -0.7422577  0.3143094 -2.3615514 1.819865e-02
age5:scarYes -0.8807367  0.3212798 -2.7413387 6.118940e-03
age6:scarYes -0.1995773  0.3457337 -0.5772572 5.637658e-01
age7:scarYes -0.6113738  0.4346772 -1.4065007 1.595755e-01
```

We can display the new estimated probabilities (in %) by age group doing:

```
grid$fit.allI <- 100*predict(e.glmAllI, newdata = grid, type = "response")
pfit.allI <- reshape(grid[,c("age","scar","fit.allI")],
                     direction = "wide", v.names = "fit.allI",
                     idvar = "age", timevar = "scar")
names(pfit.allI) <- c("age","bcg.yes","bcg.no")
pfit.allI
```

```
  age     bcg.yes     bcg.no
1   1 0.008532423 0.01316829
2   2 0.137281820 0.15376013
3   3 0.290122643 0.49654194
4   4 0.343769184 0.71942446
5   5 0.338862137 0.81366965
6   6 0.672371638 0.81967213
7   7 0.483870968 0.88813303
```

They look quite similar except for age group 4 and 5. b) To test whether there is any interaction between the vaccine effect and age we can do a likelihood ratio test:

```
anova(e.glmAllI, e.glmAll, test = "LRT")
```

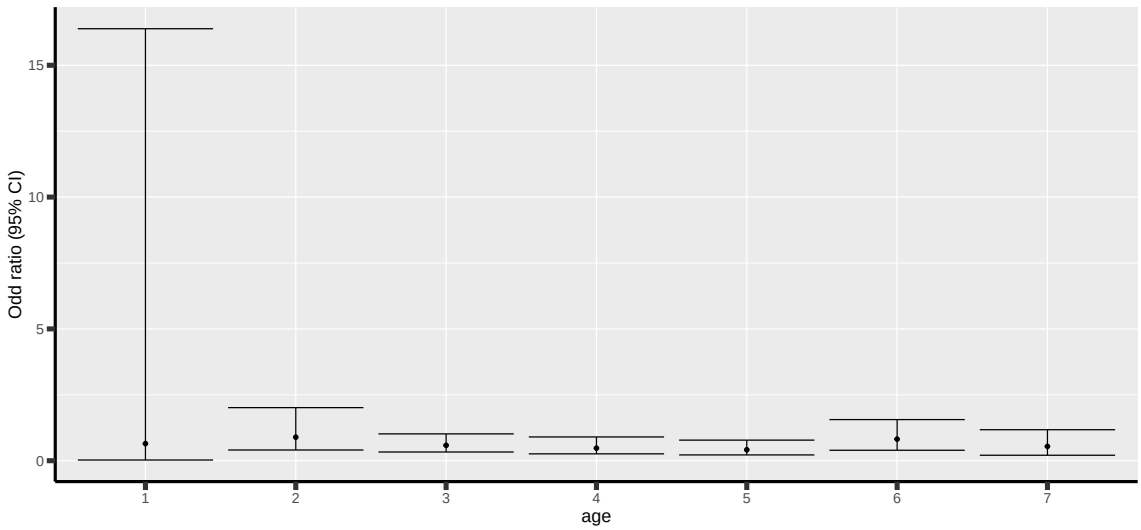
```
Analysis of Deviance Table

Model 1: status == "case" ~ age + age:scar
Model 2: status == "case" ~ age + scar
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        14     3284.4
2        20     3288.0 -6  -3.6002   0.7306
```

So our data does not provide much evidence in favor of an interaction. c) To investigate the lack of an interaction we can display the odd ratios with their (unadjusted) confidence intervals:

```
## collect all log-odds and their CIs
M.coefCI <- data.frame(estimate = coef(e.glmAllI),
                       lower = confint(e.glmAllI)[,1],
                       upper = confint(e.glmAllI)[,2])
## find those related to scar
index.scar <- grep("scar",rownames(M.coefCI))
## select these coefficient and convert to OR
M.ORCI <- cbind(age = as.character(1:7),
                exp(M.coefCI[index.scar,]))
## graphical display
library(ggplot2)
gg <- ggplot(M.ORCI, aes(x = age, y = estimate,
                         ymin = lower, ymax = upper))
gg <- gg + geom_point() + geom_errorbar()
gg <- gg + ylab("Odd ratio (95% CI)")
gg
```

```
Waiting for profiling to be done...
Waiting for profiling to be done...
```



We have very little information about the odd ratio in the first age group (it could be anything from almost 0 to 15) so we are unable to conclude about the lack of an interaction.

2. With the 1000 randomly sampled controls, we can still estimate the (log-)odd ratio, we obtain a similar value with a slightly higher standard error (loss of precision due to less controls).

```
bcg.1000 <- bcg[bcg$status %in% c("case","con1000"),]
e.glm1000 <- glm(status=="case" ~ age + scar,
                 family = binomial(link="logit"),
                 weight = n, data = bcg.1000)
summary(e.glm1000)$coef
```

```
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -4.5008399  0.7137259 -6.306118 2.861201e-10
age2         2.6684701  0.7413354  3.599545 3.187745e-04
age3         3.4711026  0.7281379  4.767095 1.869008e-06
age4         3.9232740  0.7332599  5.350455 8.773346e-08
age5         3.9700073  0.7361925  5.392621 6.943712e-08
age6         4.0451961  0.7343900  5.508240 3.624393e-08
age7         4.2062111  0.7332278  5.736568 9.661446e-09
scarYes     -0.5475497  0.1603810 -3.414056 6.400331e-04
```

However now the estimated probabilities (in %) of death are biased upward:

```
grid$fit.1000 <- 100*predict(e.glm1000, newdata = grid, type = "response")
pfit.1000 <- reshape(grid[,c("age","scar","fit.1000")],
                     direction = "wide", v.names = "fit.1000",
                     idvar = "age", timevar = "scar")
names(pfit.1000) <- c("age","bcg.yes","bcg.no")
pfit.1000
```

```
  age    bcg.yes    bcg.no
1   1  0.6378714  1.097782
2   2  8.4716802 13.795620
3   3 17.1180047 26.313504
4   4 24.5063603 35.949285
5   5 25.3812253 37.032270
6   6 26.8313980 38.801972
7   7 30.1076144 42.687103
```

3. With the matched controls, we can still estimate the (log-)odd ratio, we obtain a similar value with a slightly higher standard error (loss of precision due to less controls - theoretically lower loss compared to random sampling).

```
bcg.match <- bcg[bcg$status %in% c("case","conmatch"),]
e.glmmatch <- glm(status=="case" ~ age + scar,
                  family = binomial(link="logit"),
                  weight = n, data = bcg.match)
summary(e.glmmatch)$coef
```

```
              Estimate Std. Error     z value     Pr(>|z|)
(Intercept) -1.06665181  0.7998080 -1.33363483 0.1823235610
age2        -0.04206366  0.8270889 -0.05085748 0.9594390838
age3         0.01187032  0.8115586  0.01462657 0.9883301026
age4         0.07131384  0.8139114  0.08761867 0.9301797641
age5         0.02443657  0.8159681  0.02994795 0.9761085667
age6        -0.16276227  0.8136357 -0.20004317 0.8414468207
age7        -0.23798587  0.8128846 -0.29276711 0.7697001670
scarYes     -0.57206476  0.1546741 -3.69851782 0.0002168621
```

As before the estimated probabilities (in %) of death are biased upward but now the age effects are also biased (and non-statistically significant). But ignoring age would lead to a biased (log-)odd ratio:

```
e0.glmmatch <- glm(status=="case" ~ scar,
                  family = binomial(link="logit"),
                  weight = n, data = bcg.match)
summary(e0.glmmatch)$coef
```

```
              Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -1.1733191 0.09074598 -12.929708 3.059782e-38
scarYes     -0.4768616 0.14155193  -3.368811 7.549330e-04
```

This is to be expected as, from a theoretical point of view, when we sample the controls in an age-dependent way, we have to adjust on age in the analysis (regardless of its 'statistical significance').

4. This is because as mentionned in the `anova` output "Terms added sequentially (first to last)" so the `age` effect is tested in a model with only age and the `scar` effect in a model with only `age` and `scar` i.e. `e.glmAll` and no `e.glmAllI.bis`.