# Solution to the practicals - Logistic and Poisson regression

Epidemiological methods in medical research 2023

23 February 2023

## Exercise 1: BCG study

1. The output contains the proportion of individuals who contracted leprosy among the non-vaccinated (first line) and the vaccinated (third line) in each age group. It also displays the proportion of individual who did not contracted leprosy among the non-vaccinated (second line) and the vaccinated (fourth line) in each age group.
   By taking, for a given age, the difference or ratio between the value of the two groups we can quantify the vaccine effect. But we cannot reach a firm conclusion as we do not know the uncertainty relative to the estimation of those numbers.

2. The code perform for each age group a statistical test comparing the proportion of events between groups and collect the results into a table.
   We now have a quantifitication of the effect size (absolute risk reduction) along with the associated uncertainty. We note that we do not have enough evidence to conclude that the vaccine is beneficial to all age groups. If we where to focus on the groups with the largest benefit then we should adjust for multiple comparisons, e.g.:

```
p.adjust(df.resI$p.value, method = "bonferroni")
```

```
[1] 1.00000000 1.00000000 0.54788074 0.23284957 0.07263926 1.00000000 1.00000000
```

3. When looking at the risk ratio the vaccine effect seems more uniform in the sense that it nearly half the risk for all age groups. Note that the risk is very low in the first age group, so this benefit may not be clinically relevant.
   A nice property of this testing procedure is that the p-value for the vaccine effect is the same when considering the risk difference or the risk ratio.

[Extra] Yes we could use a binomial model with the identity link to estimate the risk difference:

```
e.glmRD <- glm(status=="case" ~ age + age:scar,
            family = binomial(link="identity"),
            weight = n,
            data = bcg.r)
summary(e.glmRD)$coef[8:14,]
```

```
                  Estimate   Std. Error    z value   Pr(>|z|)
age00_04:scar -4.635868e-05 0.0001569003 -0.2954658 0.76763808
age05_09:scar -1.647831e-04 0.0005907876 -0.2789211 0.78030537
age10_14:scar -2.064193e-03 0.0011214560 -1.8406367 0.06567482
age15_19:scar -3.756553e-03 0.0019058229 -1.9710923 0.04871332
age20_24:scar -4.748075e-03 0.0019712088 -2.4087124 0.01600891
age25_29:scar -1.473005e-03 0.0024358170 -0.6047272 0.54536024
age30_34:scar -4.042621e-03 0.0023551336 -1.7165143 0.08606795
```

and a logistic model with a log link to estimate a risk ratio

```
e.glmRR <- glm(status=="case" ~ age + age:scar,
               family = binomial(link="log"),
               weight = n,
               data = bcg.r)
cbind(RR = exp(coef(e.glmRR)), summary(e.glmRR)$coef)[8:14,]
```

```
                    RR   Estimate Std. Error    z value    Pr(>|z|)
age00_04:scar 0.6479544 -0.4339349  1.4130667 -0.3070873 0.758776900
age05_09:scar 0.8928310 -0.1133579  0.4026162 -0.2815533 0.778286037
age10_14:scar 0.5842863 -0.5373642  0.2843583 -1.8897436 0.058792257
age15_19:scar 0.4778392 -0.7384811  0.3124770 -2.3633134 0.018112343
age20_24:scar 0.4164616 -0.8759611  0.3194470 -2.7421170 0.006104459
age25_29:scar 0.8202934 -0.1980932  0.3432899 -0.5770435 0.563910048
age30_34:scar 0.5448181 -0.6073033  0.4323747 -1.4045765 0.160147263
```

The p-values differ between the two approaches as `glm` is using two different approximations to quantify the uncertainty.

4. the estimated coefficients are log-odds or log-odds ratios. This means that we should compute the exponential of the coefficients. The intercept ($\alpha$):

```
exp(coef(e.common))[1]
```

```
 (Intercept)
0.0001391389
```

gives the odds of leprosy occurence for a non-vaccinated 0-4 year old kid. The corresponding p-value is testing whether the log odds is 0, i.e. the risk is 0.5 for this age and vaccination group. The age coefficients ($\beta_1, \ldots, \beta_6$):

```
exp(coef(e.common))[2:7]
```

```
age05_09 age10_14 age15_19 age20_24 age25_29 age30_34
13.78438 35.98533 45.79287 49.41018 63.79227 63.92043
```

gives the odds ratio for the age effect. They suggest an increase risk of leprosy when getting older. The scar coefficient ($\gamma$):

```
exp(coef(e.common))[8]
```

```
     scar
0.5783652
```

gives the odd ratio for the vaccine effect. It suggests that when considering individuals from the same age group, those vaccinated have a reduced risk of leprosy compared to the non-vaccinated[1]. Since the disease is rare, this odd ratio is a good approximation of the risk ratio (which vary between 0.57 and 0.58 according to this model). So it is reasonnable to communicate an estimate risk ratio of 0.57 with its confidence interval:.

```
exp(confint(e.common)["scar",])
```

```
Waiting for profiling to be done...
    2.5 %    97.5 %
0.4382287 0.7617475
```

5. The second logistic model is a re-parametrisation of the first where the log-odds of the disease for each age group among the non-vaccinated is displayed instead of the log-odds for a reference group and log-odds ratios between groups. The estimated vaccine effect is identical.

6. The first predicted value is the log-odds for each age and vaccination sub-group. The second is the estimated probability of leprosy. They would be the same for the other parametrisation.
   Both can be computed based on the estimated coefficients, e.g. see the tables in appendix B for the probabilities. One can also use the design matrix and perfom matrix operations:

[1] ⚠ a causal interpretation like "there is evidence that the vaccine reduces the risk of leprosy" is only valid upon certain assumptions like no unmeasured confounders. A non-causal interpretation "the risk is lower in this group" does not require such assumption but does not help to decide about the vaccine efficacy. It can only be used for predictive purpose, e.g. if you belong to the vaccinated group you can expect to be "protected" but that may not due to the vaccine but to other things (e.g. better access to healthcare).

```
X <- model.matrix(e.common)[1:5,] ## design matrix
X
```

| | (Intercept) | age05_09 | age10_14 | age15_19 | age20_24 | age25_29 | age30_34 | scar |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 5 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

```
beta <- coef(e.common)
beta
```

```
(Intercept)    age05_09    age10_14    age15_19    age20_24    age25_29
 -8.8800380   2.6235363   3.5831114   3.8241284   3.9001565   4.1556320
   age30_34        scar
  4.1576390  -0.5470646
```

```
cbind(X %*% beta, ## log-odds
      1/(1+exp(-X %*% beta)) ## probability
      )
```

```
       [,1]         [,2]
1 -9.427103 8.050566e-05
2 -6.803566 1.108580e-03
3 -5.843991 2.888886e-03
4 -5.602974 3.673339e-03
5 -5.526946 3.962357e-03
```

7. We can compare the estimated probabilities under this model to the unrestricted model of Part I. Major discrepancies would indicate lack of fit. We should keep in mind that both are estimated with uncertainty so discrepancies when the uncertainty is high are not as concerning as when the uncertainty is low.

[Extra] Because the risk difference is very variable between age groups. Since the risk is very close to 0 in the first age group, considering a large risk difference between age groups would lead to negative probabilites and therefore issues when computing the log-likelihood. This is why it is important for the software to have good starting value: it needs to be able to evaluate the log-likelihood and its derivative to find the values that best fit the data.

8. The estimated coefficients are again log-odds or log-odds ratios. This means that we should once more compute the exponential of the coefficients. The intercept ($\alpha$):

```
exp(coef(e.full))[1]
```

```
 (Intercept)
0.0001317003
```

gives the odds of leprosy occurence for a non-vaccinated 0-4 year old kid, as before. The corresponding p-value is testing whether the log odds is 0, i.e. the risk is 0.5 for this age and vaccination group. The age coefficients $(\beta_1, \ldots, \beta_6)$:

```
exp(coef(e.full))[2:7]
```

```
age05_09 age10_14 age15_19 age20_24 age25_29 age30_34
11.69299 37.89057 55.02174 62.28876 62.75207 68.04023
```

gives the odds ratio for the age effect, as before. They suggest an increase risk of leprosy when getting older. The scar coefficient ($\gamma$):

```
exp(coef(e.full))[8]
```

```
     scar
0.6479222
```

gives the odd ratio for the vaccine effect among the 0-4 year old kids. Finally the interaction coefficients ($\delta_1, \ldots, \delta_6$):

```
exp(coef(e.full))[9:14]
```

```
age05_09:scar age10_14:scar age15_19:scar age20_24:scar age25_29:scar
    1.3777638     0.8999178     0.7347147     0.6397025     1.2641594
age30_34:scar
    0.8374538
```

gives the odds ratio for the change in vaccine effect across age. For instance when considering the 20-24 year old, the vaccine seems to offer more protection. ⚠ This coefficient does not (only) compare the risk between vaccinated 20-24 and non-vaccinated 20-24! It compares the odds ratio of the 20-24 (vaccinated vs. not vaccinated) to the odds ratio of the 0-4 (vaccinated vs. not vaccinated). To quantify the vaccination effect, we report the odds ratio among the 0-4, as well as the odds ratios in the other age categories:

```
c(exp(coef(e.full))[8],
  exp(coef(e.full)[8] + coef(e.full)[9:14]))
```

```
         scar age05_09:scar age10_14:scar age15_19:scar age20_24:scar
    0.6479222     0.8926837     0.5830767     0.4760379     0.4144775
age25_29:scar age30_34:scar
    0.8190769     0.5426049
```

Ignoring uncertainty, we see a beneficial effect in every age group.
The predicted probabilities are the same as in part I:

```
pred.full <- data.frame(age = bcg.r$age[1:14],
                        scar = bcg.r$scar[1:14],
                        pred1 = predict(e.full, type = "link")[1:14],
                        pred2 = predict(e.full, type = "response")[1:14])
head(pred.full)
```

```
    age scar     pred1        pred2
1 00_04    1 -9.368967 8.532423e-05
2 05_09    1 -6.589516 1.372818e-03
3 10_14    1 -5.839716 2.901226e-03
4 15_19    1 -5.669511 3.437692e-03
5 20_24    1 -5.683938 3.388621e-03
6 25_29    1 -4.995368 6.723716e-03
```

9. This new model is a reformulation of the previous model showing the log-odds of the disease in each age group and the log-odds ratio of the vaccine effect for each group.

10. In the first `anova`, the last F-test (`age:scar`) is testing whether the vaccine effect varies across age groups. The remainder should not be considered as they are for models without interaction. ⚠ Absence of evidence is not evidence of absence. No evidence for heterogeneity does not mean that the risk reduction is the same in each age group (we may just not have enough statistical power to detect it).
    In the second `anova`, the last F-test (`age:scar`) is testing whether there is any effect of the vaccine at any age. Note that doing a global test for the vaccination effect takes care multiple comparison issues but is not very helpful to understand what is going on. Testing separately the vaccination effect in each age group will give more inside. We can see that (ignoring uncertainty) the risk reduction varies across age groups and it is the largest for 20-24.

11. Part I describes a fully stratified model on the probability scale:

    ✔ it is easy to understand

    ✔ few assumptions

    ✔ confidence intervals and p-values can be estimated accuratly even with small sample

    ✘ it is limited to categorical covariates

    ✘ it leads to multiple comparisons which cannot (easily) be accounted for in an efficient way.

    ✘ can be lengthy to communicate as results are age-specific

Part II describes a common effect model on the log-odd scale. It assumes a constant vaccine effect over age groups on the log odd scale.

✔ no multiple comparison issue as only a single coefficient for the vaccine effect is estimated.

✘ modeling is performed on the log-odd scale which can be challenging to understand and communicate. However in this study it is possible to move back to the probability scale.

✘ this 'common effect' assumption should be checked.

✔ it can handle categorical and continuous covariates

✘ 'default' confidence intervals and p-values are not very accurate in small samples

Part III describes a fully stratified model on the log-odd scale. It is similar to Part I but implemented in a different way:

✘ modeling is performed on the log-odd scale which can be challenging to understand and communicate. However in this study it is possible to move back to the probability scale.

✔ few assumptions

✔ multiple comparisons can be handled via a F-test.

✔ it can handle categorical and continuous covariates

✘ 'default' confidence intervals and p-values are not very accurate in small samples

# Exercise 2: The Bissau study

**1)** We fit a poisson model with BCG vaccine as covariate and output the estimated rate/rate ratios:

```
mB <- glm(fupstatus=="dead" ~ bcg + factor(agem),
          family = poisson(link = "log"),
          offset = log(fuptime),
          data = bissau)

exp(cbind(coef(mB), confint(mB)))
```

```
Waiting for profiling to be done...
                                 2.5 %       97.5 %
(Intercept)   0.000277131 0.0001961155 0.0003785185
bcgyes        0.708006220 0.5326071349 0.9447067024
factor(agem)1 1.122316129 0.7114309104 1.7733213337
factor(agem)2 0.774247396 0.4623273571 1.2802498059
factor(agem)3 1.219971430 0.7554826439 1.9673791675
factor(agem)4 1.393708253 0.8663140583 2.2430314247
factor(agem)5 1.390276273 0.8495469068 2.2679539181
factor(agem)6 0.984781408 0.4706214535 1.9051326578
```

or equivalently using the Epi package:

```
library(Epi)
mB2 <- glm(cbind(fupstatus=="dead", fuptime) ~ bcg + factor(agem),
          family = poisreg,
          data = bissau)

ci.exp(mB2)
```

```
               exp(Est.)        2.5%      97.5%
(Intercept)   0.000277131 0.000199661 0.00038466
bcgyes        0.708006222 0.531780215 0.94263155
factor(agem)1 1.122316129 0.712183864 1.76863526
factor(agem)2 0.774247416 0.466415144 1.28524785
factor(agem)3 1.219971428 0.757358499 1.96515955
factor(agem)4 1.393708251 0.867612350 2.23881401
factor(agem)5 1.390276271 0.852446540 2.26743616
factor(agem)6 0.984781409 0.492749924 1.96812699
```

We see that the children with a BCG vaccination has a 30% lower mortality.

**2)** We now fit a logistic model with DTP as covariate:

```
mD <- glm(cbind(fupstatus=="dead",fuptime) ~ dtpany + factor(agem),
          family = poisreg,
          data = bissau)

ci.exp(mD)
```

```
                exp(Est.)         2.5%        97.5%
(Intercept)   0.000254857 0.0001846531 0.0003517518
dtpanyTRUE    1.003186792 0.7235717705 1.3908554490
factor(agem)1 1.030067175 0.6563530919 1.6165664445
factor(agem)2 0.679706693 0.4074048320 1.1340100844
factor(agem)3 1.039418102 0.6328425781 1.7072018030
factor(agem)4 1.172477180 0.7101992364 1.9356578649
factor(agem)5 1.157295393 0.6875997634 1.9478375342
factor(agem)6 0.808277086 0.3941463858 1.6575360627
```

Here we don't see evidence for an effect of DTP vaccine on mortality.

**3)** We fit a poisson model with age, BCG, and DTP as covariates:

```
mBD <- glm(cbind(fupstatus=="dead",fuptime) ~ bcg + dtpany + factor(agem),
          family = poisreg,
          data = bissau)

ci.exp(mBD)
```

```
                 exp(Est.)         2.5%        97.5%
(Intercept)   0.0002878654 0.0002071296 0.0004000706
bcgyes        0.5763750973 0.3937288248 0.8437488745
dtpanyTRUE    1.4467080915 0.9464221379 2.2114490123
factor(agem)1 1.1426491979 0.7247307695 1.8015616894
factor(agem)2 0.7309060595 0.4370156175 1.2224361016
factor(agem)3 1.0976914964 0.6677740973 1.8043925723
factor(agem)4 1.2279150216 0.7439773842 2.0266413098
factor(agem)5 1.2076342634 0.7178336764 2.0316412591
factor(agem)6 0.8454635468 0.4124556564 1.7330556581
```

We see a slightly higher effect of BCG and a lower mortality among those without DTP. We can make a table of events and person-years for the two types of vaccines:

```
tBD <- xtabs( cbind("death" = fupstatus=="dead",
                    "person-year" = fuptime/365.25) ~ bcg+dtpany,
             data = bissau)
ftable(tBD)
```

```
                death person-year
bcg dtpany
no  FALSE    95.00000   875.61123
    TRUE      2.00000    14.89665
yes FALSE    33.00000   537.59890
    TRUE     92.00000   981.71389
```

We see that the number of person-years among those with DTP but no BCG is tiny (14.9 person years). So the effect of DTP is essentially only interpretable as the effect of DTP within those with a BCG.

**4)** We saw in the previous table that we had very little information about DTP without BCG. So should not be able to decide whether there is an interaction effect.

```
mI <- glm(cbind(fupstatus=="dead",fuptime) ~ bcg * dtpany + factor(agem),
          family = poisreg,
          data = bissau)
ci.exp(mI)
```

```
                     exp(Est.)         2.5%        97.5%
(Intercept)       0.0002887031 0.0002076286 0.0004014356
bcgyes            0.5667796712 0.3796763080 0.8460870190
dtpanyTRUE        1.1982140799 0.2912669092 4.9292141869
factor(agem)1     1.1457782896 0.7264444454 1.8071690097
factor(agem)2     0.7325344936 0.4378941837 1.2254256949
factor(agem)3     1.0987117546 0.6683699179 1.8061368222
factor(agem)4     1.2291112122 0.7446953489 2.0286340907
factor(agem)5     1.2105119947 0.7194290387 2.0368086503
factor(agem)6     0.8467886100 0.4130882576 1.7358299027
bcgyes:dtpanyTRUE 1.2277034217 0.2858619043 5.2726707164
```

This is confirmed by the output of the poisson regression where the confidence interval for the interaction extends from 0.3 to 5, i.e. is very large. So we cannot rule out no effect, a protective effect, or a harmful effect, i.e. we don't really know.