Solution to the practicals - Dealing with confounding: DAGs and stratification

Epidemiological methods in medical research 2023

2 February 2023

Exercise 1: warming-up with DAGs

1. We should not adjust on any variable:

   - adjusting on stroke would bias our estimate as we would only consider the direct effect instead of the total effect.
   - adjusting on obesity would (slightly) decrease the efficiency of the estimator (i.e. lead to more sampling error) as it is not related to the outcome.

2. In this new DAG, obesity is confouding the relationship between hypertension and death. For instance it can be that obese persons are more likely to have hypertension while obesity also increases the risk of death (independently of hypertension). Ignoring obesity, we would not know whether a difference in mortality would be due to hypertension or to obesity. We should adjust (only) on obesity.
   When including both obesity and hypertension in a model, we will not be able to estimate the total effect of obesity on the risk of death. Indeed, the indirect effect through hypertension is blocked so we would only estimate the direct effect.

- In DAG (c) there is a direct path between hypertension and death and two indirect paths:

  - a causal path passing through stroke. It is open if we don't adjust on stroke.
  - a non causal path passing through Diet-Obesity-Gene. It is closed if we don't adjust on obesity as obesity is a collider.

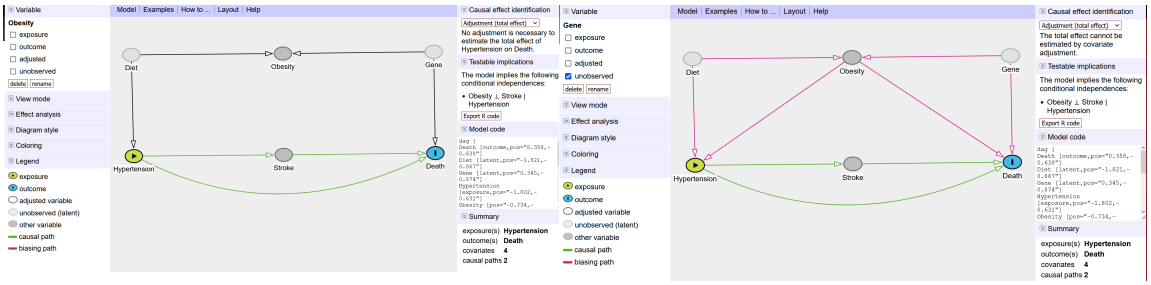
  So we can estimate the causal effect by not adjusting on any variable.

DAG (d) is more difficult as there are now three indirect paths:

- a causal path passing through stroke. It is open if we don't adjust on stroke.
- a non causal path passing through obesity. It can be closed by adjusting on obesity as obesity is a confounder.
- a non causal path passing through diet-obesity-gene. Adjusting on obesity open this path so we would need to adjust on Gene or Diet to close it (as they are "confounders").

So we cannot estimate the causal effect as we don't measure diet or gene.

This is in line with what DAGitty outputs:



Exercise 2: Analysis of the UC Berkeley data

1. There are several ways to compute the percentage of admission. The easier may be to use the "table" format:

```
c(
sum(UCBAdmissions["Admitted","Male",])/sum(UCBAdmissions[,"Male",]),
sum(UCBAdmissions["Admitted","Female",])/sum(UCBAdmissions[,"Female",])
)
```

```
[1] 0.4451877 0.3035422
```

Otherwise one can agreggate numbers across departments and then compute the probability of admission by gender:

```
tableGlobal <- xtabs(cbind(N,D) ~ Gender, data = df)
tableGlobal[,"D"]/tableGlobal[,"N"]
```

```
     Male    Female
0.4451877 0.3035422
```

2. We can perform a Cochran-Mantel-Haenszel (CMH) test using:

```
mantelhaen.test(UCBAdmissions[,c("Female","Male"),])
```

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

data:  UCBAdmissions[, c("Female", "Male"), ]
Mantel-Haenszel X-squared = 1.4269, df = 1, p-value = 0.2323
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.9431028 1.2954922
sample estimates:
common odds ratio
         1.105343
```

The common odd ratio can be computed using the formula from the lecture:

```
a <- UCBAdmissions["Rejected","Male",]
b <- UCBAdmissions["Admitted","Male",]
c <- UCBAdmissions["Rejected","Female",]
d <- UCBAdmissions["Admitted","Female",]
n <- a + b + c + d
sum(a*d/n)/sum(b*c/n)
```

```
[1] 1.105343
```

When computing a common odd ratio, we assume that the gender effect is similar across all departments. To check that we can compute the odd ratio in each department:

```
OR <- a*d/(b*c)
OR
```

```
        A         B         C         D         E         F
2.8635896 1.2461048 0.8825661 1.0854419 0.8185776 1.2079151
```

and see only a large odd ratio in department A, otherwise the odd ratio is close to 1 and sometimes smaller. This hetetogeneity is confirmed by the Breslow-Day test:

```
DescTools::BreslowDayTest(UCBAdmissions)
```

```
        Breslow-Day test on Homogeneity of Odds Ratios

data:  UCBAdmissions
X-squared = 18.826, df = 5, p-value = 0.002071
```

So assuming a common odd ratio does not seem reasonnable and we should find another test than the CMH test. The statistical reason is that is won't be a very powerful test since the common odd ratio assumption is not fulfield. The intuitive reason is that since the gender effect appears mostly in one department, it will get diluted when pooling over departments thus giving low power to reject the null. We get a very similar result when estimating the common odd ratio via a logistic model:

```
df$Gender <- relevel(df$Gender, "Male")
e.common <- glm(cbind(D,N-D) ~ Dept + Gender,
                data = df, family = binomial(link="logit"))
res.common <- summary(e.common)$coef
res.common <- cbind(res.common,
                    OR = exp(res.common[,"Estimate"]))
res.common["GenderFemale",,drop=FALSE]
```

```
              Estimate Std. Error  z value  Pr(>|z|)       OR
GenderFemale 0.09987009 0.08084647 1.235306 0.2167168 1.105027
```

In either case we do not have compelling evidence for a difference in percentage of admission between male and female:

⚠ that does necessarily imply that we have evidence for no gender difference.

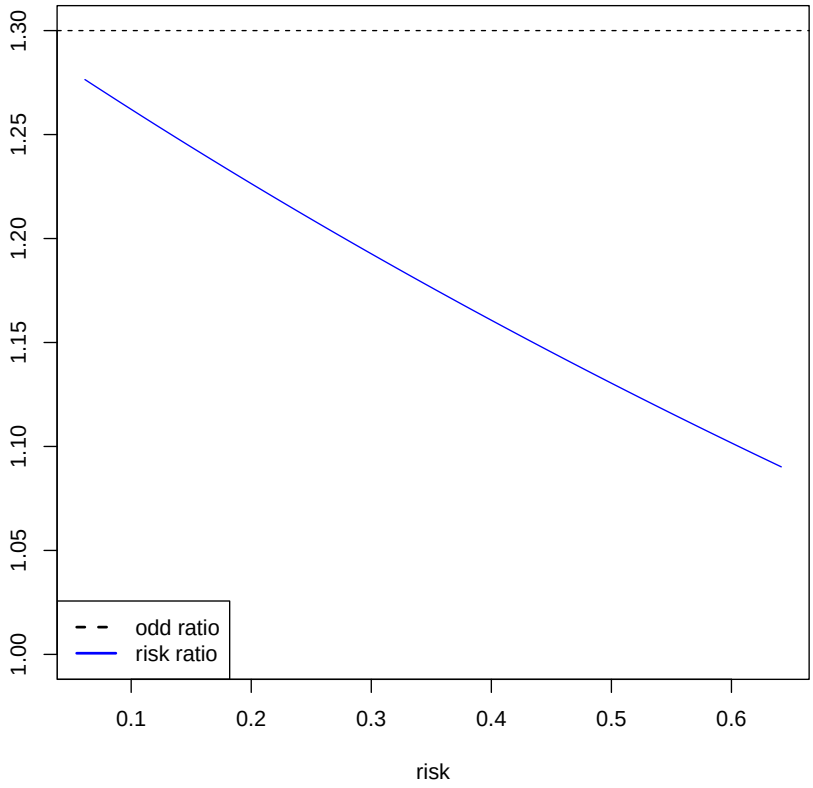
Looking at the confidence interval of the CMH OR, the data is compatible with a OR as large as 1.3. Unfortunately the OR scale is not intuitive so we will use the logistic regression to move back to the probability scale, i.e. output the fitted admission probabilities in each department and for each gender:

```
grid <- df[,c("Dept","Gender")]
grid$fit.common <- 100*predict(e.common,
                               newdata = grid, type = "response")
## move to wide format (i.e. one row per Department)
reshape(grid, direction = "wide", v.names = "fit.common",
        idvar = "Dept", timevar = "Gender")
```

```
   Dept fit.common.Male fit.common.Female
1     A       64.153930          66.41674
3     B       63.149912          65.44196
5     C       33.613931          35.87769
7     D       32.903451          35.14470
9     E       23.916654          25.78096
11    F        6.154717           6.75745
```

So the frequency of admission varies between 0.06 and 0.64 in the reference group. Applying the formula to move from an odd ratio to a risk ratio, we get:

```
seqRisk <- seq(6.15/100,64.15/100, length.out = 100)
OR <- 1.3
plot(seqRisk,OR/(1-seqRisk+seqRisk*OR), type = "l", col = "blue",
     ylim = c(1,1.3), xlab = "risk", ylab = "")
abline(h = OR, lty = 2)
legend("bottomleft", legend = c("odd ratio","risk ratio"),
       lwd = 2, lty = c(2,1), col = c("black","blue"))
```



i.e. the risk ratio could be as large as 10-25% depending on the department. This is not really neglectable so we would need to gather more data to conclude about the absence of gender difference.

Note: one advantage of the Mantel-Haenszel odd ratio is that we can explicit the contribution of each department:

```
weight <- b*c/n
100*weight/sum(weight)
```

```
       A         B         C         D         E         F
7.152125  3.311327 35.059726 29.163317 18.613473  6.700033
```

as the common odd ratio is a weighted average of the previous weights

```
weighted.mean(OR, w = weight)
```

```
[1] 1.105343
```

3. We first fit a logistic regression assuming no gender effect but modeling a department effect:

```
e.H0 <- glm(cbind(D,N-D) ~ Dept,
            data = df, family = binomial(link="logit"))
summary(e.H0)$coef
logLik(e.H0)
```

```
               Estimate Std. Error     z value     Pr(>|z|)
(Intercept)  0.59345997 0.06838086   8.6787441 4.001629e-18
DeptB       -0.05059499 0.10968048  -0.4612944 6.445874e-01
DeptC       -1.20914909 0.09725937 -12.4322124 1.747305e-35
DeptD       -1.25833005 0.10151581 -12.3954096 2.767280e-35
DeptE       -1.68296057 0.11733283 -14.3434759 1.170434e-46
DeptF       -3.26910674 0.16706908 -19.5673951 2.932940e-85
'log Lik.' -45.3376 (df=6)
```

The interpretation of the coefficient is not completely straightforward. They are log odds and log odd ratios - see appendix for more details. We then fit a model accounting for a different gender effect in each department:

```
e.H1 <- glm(cbind(D,N-D) ~ Dept*Gender,
            data = df, family = binomial(link="logit"))
summary(e.H1)$coef
logLik(e.H1)
```

```
                     Estimate Std. Error     z value     Pr(>|z|)
(Intercept)         0.49212143 0.07174966   6.8588682 6.940825e-12
DeptB               0.04162783 0.11318919   0.3677721 7.130431e-01
DeptC              -1.02763967 0.13549685  -7.5842331 3.344593e-14
DeptD              -1.19607953 0.12640656  -9.4621632 3.016374e-21
DeptE              -1.44908321 0.17681152  -8.1956378 2.492678e-16
DeptF              -3.26186520 0.23119594 -14.1086615 3.358928e-45
GenderFemale        1.05207596 0.26270810   4.0047336 6.208742e-05
DeptB:GenderFemale -0.83205342 0.51039480  -1.6302153 1.030560e-01
DeptC:GenderFemale -1.17699758 0.29955796  -3.9291147 8.525915e-05
DeptD:GenderFemale -0.97008876 0.30261874  -3.2056467 1.347593e-03
DeptE:GenderFemale -1.25226298 0.33032201  -3.7910371 1.500195e-04
DeptF:GenderFemale -0.86318013 0.40266653  -2.1436600 3.206014e-02
'log Lik.' -34.46984 (df=12)
```

The coefficients are also log odd and log odd ratios - see again appendix for more details. We can now perform the likelihood ratio test by comparing the likelihood between the two models:

```
anova(e.H0, e.H1, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: cbind(D, N - D) ~ Dept
Model 2: cbind(D, N - D) ~ Dept * Gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         6     21.735
2         0      0.000  6   21.735 0.001352 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is evidence for a gender effect. However the likelihood ratio test does not help much in knowning the direction or magnitude of the effect.

4. There are several ways to compute the percentage of admission per gender. One can use the "table" format:

```
tablePC <- UCBAdmissions["Admitted",,]/ (UCBAdmissions["Rejected",,] +
   UCBAdmissions["Admitted",,])
round(100*tablePC,2)
```

```
        Dept
Gender       A     B     C     D     E     F
  Male   62.06 63.04 36.92 33.09 27.75  5.90
  Female 82.41 68.00 34.06 34.93 23.92  7.04
```

or the "data.frame" format:

```
cbind(df,pc=round(100*df$D/df$N,2))
```

```
   Gender Dept   N   D    pc
1    Male    A 825 512 62.06
2  Female    A 108  89 82.41
3    Male    B 560 353 63.04
4  Female    B  25  17 68.00
5    Male    C 325 120 36.92
6  Female    C 593 202 34.06
7    Male    D 417 138 33.09
8  Female    D 375 131 34.93
9    Male    E 191  53 27.75
10 Female    E 393  94 23.92
11   Male    F 373  22  5.90
12 Female    F 341  24  7.04
```

We can estimate the difference in probabilities using:

```
tablePC["Female",]-tablePC["Male",]
```

```
         A           B           C           D           E           F
0.20346801  0.04964286 -0.02858996  0.01839808 -0.03830116  0.01140000
```

P-values can be obtained using:

```
exact2x2::uncondExact2x2(x1 = UCBAdmissions["Admitted","Female","A"],
                         n1 = sum(UCBAdmissions[,"Female","A"]),
                         x2 = UCBAdmissions["Admitted","Male","A"],
                         n2 = sum(UCBAdmissions[,"Male","A"]),
                         parmtype = "difference",
                         conf.int = TRUE)
```

```
        Unconditional Exact Test on Difference in Proportions, method=
        FisherAdj, central

data:  x1/n1=(89/108) and x2/n2= (512/825)
proportion 1 = 0.82407, proportion 2 = 0.62061, p-value = 1.442e-05
alternative hypothesis: true p2-p1 is not equal to 0
95 percent confidence interval:
 -0.31567955 -0.02065659
sample estimates:
    p2-p1
-0.203468
```

Even after adjustment for 6 comparisons this p-value stays far below 0.05 meaning that there is evidence of a gender effect. In particular the admission rate is higher for females in Department A.

Appendix A: Additional R code

A.1 Exercice 3, question 3a (interpretation of the coefficients)

In the logistic model under the null, the intercept is the log odd of admission, i.e. the probability of admission in department A is:

```
odd <- exp(coef(e.H0)["(Intercept)"])
odd/(1+odd)
```

```
(Intercept)
  0.6441586
```

while the other coefficients are the log odd ratio for the probability admission in one department vs. department A. The probability of admission in each department can be obtained by:

```
beta <- exp(coef(e.H0)[paste0("Dept", c("B","C","D","E","F"))])
odd*beta/(1+odd*beta)
```

```
     DeptB      DeptC      DeptD      DeptE      DeptF
0.63247863 0.35076253 0.33964646 0.25171233 0.06442577
```

A.2 Exercice 3, question 3b (interpretation of the coefficients)

In the logistic model under the alternative, the interpretation of the first six coefficients is the same as under the null except that it is only valid for the males, i.e. the probability of admission in department A for the males is:

```
odd <- exp(coef(e.H1)["(Intercept)"])
odd/(1+odd)
```

```
(Intercept)
  0.6206061
```

and for the other departments:

```
beta <- exp(coef(e.H1)[paste0("Dept", c("B","C","D","E","F"))])
odd*beta/(1+odd*beta)
```

```
     DeptB      DeptC      DeptD      DeptE      DeptF
0.63035714 0.36923077 0.33093525 0.27748691 0.05898123
```

The coefficient `GenderFemale` is the odd ratio for the female effect common to all strata while the rest of the coefficients are the odd ratio for the female effect in a given department vs. department A. So in department A, the probability of admission for the females is

```
gamma <- exp(coef(e.H1)["GenderFemale"])
odd*gamma/(1+odd*gamma)
```

```
(Intercept)
  0.8240741
```

and in the other departmentss it is:

```
coef.delta <- paste0("Dept",c("B","C","D","E","F"),":GenderFemale")
delta <- exp(coef(e.H1)[coef.delta])
odd*gamma*delta*beta/(1 + odd*gamma*delta*beta)
```

```
DeptB:GenderFemale DeptC:GenderFemale DeptD:GenderFemale DeptE:GenderFemale
        0.68000000         0.34064081         0.34933333         0.23918575
DeptF:GenderFemale
        0.07038123
```

A.3 Exercice 3, question 3 "Extra time"

We can do the test "manually":

- first computing the difference in likelihood:

```
diffLik <- logLik(e.H1)-logLik(e.H0)
diffLik
```

```
'log Lik.' 10.86775 (df=12)
```

and then compute the p-value based on the cumulative distribution function of the chi-squared distribution with 6 degrees of freedom:

```
1-pchisq(2*diffLik, df = 6)
```

```
'log Lik.' 0.001351993 (df=12)
```

The log-likelihood can be compute based on the number of each event (admission, rejection) and the modeled probabilities:

```
logLik.H0 <- sum((b+d)*log((b+d)/n)+(a+c)*log((a+c)/n))
logLik.H1 <- sum(b*log(b/(a+b))+a*log(a/(a+b))+c*log(c/(c+d))+d*log(d/(c+d
    )))
logLik.H1 - logLik.H0
```

```
[1] 10.86775
```

A.4 Exercice 3, question 4 "Extra time"

The "naive" estimates of the probability of admission under no gender effect are:

```
(b+d)/n
```

```
         A          B          C          D          E          F
0.64415863 0.63247863 0.35076253 0.33964646 0.25171233 0.06442577
```

These values are the same as those "predicted" by the logistic model:

```
cbind(df, fit = round(100*predict(e.H0, type = "response"),2))
```

```
   Gender Dept   N   D   fit
1    Male    A 825 512 64.42
2  Female    A 108  89 64.42
3    Male    B 560 353 63.25
4  Female    B  25  17 63.25
5    Male    C 325 120 35.08
6  Female    C 593 202 35.08
7    Male    D 417 138 33.96
8  Female    D 375 131 33.96
9    Male    E 191  53 25.17
10 Female    E 393  94 25.17
11   Male    F 373  22  6.44
12 Female    F 341  24  6.44
```

We already computed the "naive" estimates of the probablity of admission assuming a department-specific gender effect:

```
cbind(df,pc=round(100*df$D/df$N,2))
```

```
   Gender Dept   N   D    pc
1    Male    A 825 512 62.06
2  Female    A 108  89 82.41
3    Male    B 560 353 63.04
4  Female    B  25  17 68.00
5    Male    C 325 120 36.92
6  Female    C 593 202 34.06
7    Male    D 417 138 33.09
8  Female    D 375 131 34.93
9    Male    E 191  53 27.75
10 Female    E 393  94 23.92
11   Male    F 373  22  5.90
12 Female    F 341  24  7.04
```

These values are the same as those "predicted" by the logistic model:

```
cbind(df, fit = round(100*predict(e.H1, type = "response"),2))
```

```
   Gender Dept   N   D   fit
1    Male    A 825 512 62.06
2  Female    A 108  89 82.41
3    Male    B 560 353 63.04
4  Female    B  25  17 68.00
5    Male    C 325 120 36.92
6  Female    C 593 202 34.06
7    Male    D 417 138 33.09
8  Female    D 375 131 34.93
9    Male    E 191  53 27.75
10 Female    E 393  94 23.92
11   Male    F 373  22  5.90
12 Female    F 341  24  7.04
```