Week 9 - Exercises-Solutions

Exercise solutions

Stata code and output

You can certainly remember that we can derive an OR from a 2x2 table. Using the same WCGS data, carry out the following analysis:

a) Reproduce the exploratory analysis with the χ^2-test

```
use wcgs.dta
tabulate chd69 arcus, col row chi2
disp (102*2058)/(153*839)
## . use wcgs.. tabulate chd69 arcus, col row chi2
##
## +-------------------+
## | Key               |
## |-------------------|
## |     frequency     |
## |  row percentage   |
## | column percentage |
## +-------------------+
##
##            |         arcus
##      chd69 |         0          1 |     Total
## -----------+----------------------+----------
##          0 |     2,058        839 |     2,897
##            |     71.04      28.96 |    100.00
##            |     93.08      89.16 |     91.91
## -----------+----------------------+----------
##          1 |       153        102 |       255
##            |     60.00      40.00 |    100.00
##            |      6.92      10.84 |      8.09
## -----------+----------------------+----------
##      Total |     2,211        941 |     3,152
##            |     70.15      29.85 |    100.00
##            |    100.00     100.00 |    100.00
##
```

```
##           Pearson chi2(1) =  13.6382   Pr = 0.000
##
## . disp (102*2058)/(153*839)
## 1.6352801
```

There is a clear association between arcus and CHD with Chi2=13.64 and a p-value smaller than 0.001.

b) Compute the OR and check that is exactly the same result as the one obtained via simple logistic regression

```
use wcgs.dta
disp (102*2058)/(153*839)
logistic chd69 arcus
## . use wcgs.. disp (102*2058)/(153*839)
## 1.6352801
##
## . logistic chd69 arcus
##
## Logistic regression                                   Number of obs =  3,152
##                                                       LR chi2(1)    =  12.98
##                                                       Prob > chi2   = 0.0003
## Log likelihood = -879.10783                           Pseudo R2     = 0.0073
##
## ------------------------------------------------------------------------------
##        chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##        arcus |   1.635281   .2195036     3.66   0.000     1.257001    2.127399
##        _cons |    .074344   .0062298   -31.02   0.000     .0630839    .0876141
## ------------------------------------------------------------------------------
## Note: _cons estimates baseline odds.
```

A manual calculation returns OR=1.635 which is exactly the point estimate provided by logistic regression in an unadjusted analysis, OR-1.635, 95%CI=(1.26 ; 2.13)

c) A large sample formula for the standard error of the log-OR estimate in a 2x2 table is given by: $SE(log(\hat{OR})) = \sqrt{1/a + 1/b + 1/c + 1/d)}$ where a, b, c and d are the frequencies in the 2x2 table. Compute the 95% CI for the estimate you have just computed. How does it compare with the 95% obtained from logistic regression. Hint: start by computing a 95% CI for the log-OR.

```
scalar OR=(102*2058)/(153*839)
disp  OR
scalar SElog=sqrt(1/2058 + 1/839 + 1/153 + 1/102)
scalar lower =log(OR)-1.96*SElog
scalar upper =log(OR)+1.96*SElog
disp exp(lower)
disp exp(upper)
## . scalar OR=(102*2058)/(153*839. disp  OR
## 1.6352801
##
## . scalar SElog=sqrt(1/2058 + 1/839 + 1/153 + 1/102)
##
## . scalar lower =log(OR)-1.96*SElog
##
## . scalar upper =log(OR)+1.96*SElog
##
## . disp exp(lower)
## 1.2569944
##
## . disp exp(upper)
## 2.1274089
```

A similar 95% CI is obtained. Note that in general you may see a small difference since SEs are computed using Woolf's formula. The information matrix is used to compute SEs in logistic regression.

R code and output

a) Reproduce the exploratory analysis with the χ^2-test

```
wcgs <- read.csv("https://www.dropbox.com/s/uc29ddv337zcxk6/wcgs.csv?dl=1")
wcgs$arcus <- factor(wcgs$arcus)
table(wcgs$chd69,wcgs$arcus)
##
##        0    1
##   0 2058  839
##   1  153  102
chisq.test(wcgs$chd69,
           wcgs$arcus,
           correct=FALSE)
##
##  Pearson's Chi-squared test
```

```
##
## data:  wcgs$chd69 and wcgs$arcus
## X-squared = 13.638, df = 1, p-value = 0.0002216

#Alternative: Using the gtsummary library to create a table
wcgs %>%                              #data
  select(arcus, chd69) %>%            #selects the vars for table
  tbl_summary( by=chd69 ) %>%         #describes arcus by chd69
  add_p()                             #adds pvalue
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,897	**1**, N = 257	**p-value**
arcus			<0.001
0	2,058 (71%)	153 (60%)	
1	839 (29%)	102 (40%)	
Unknown	0	2	

There is a clear association between arcus and CHD with Chi2=13.64 and a p-value p=0.0002. Note that the default R function add a continuity correction (suppressed here with the option *Correct=FALSE*). Of no practical significance in large samples.

b) Compute the OR and check that is exactly the same result as the one obtained via simple logistic regression

```
OR <- 102*2058/(153*839)                  #manual calculation of the OR
OR
## [1] 1.63528

model0<-glm(chd69 ~ arcus,  family=binomial, data=wcgs)
summary(model0)
##
## Call:
## glm(formula = chd69 ~ arcus, family = binomial, data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4790  -0.4790  -0.3787  -0.3787   2.3112
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5991     0.0838 -31.016  < 2e-16 ***
## arcus1        0.4918     0.1342   3.664 0.000248 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1771.2  on 3151  degrees of freedom
## Residual deviance: 1758.2  on 3150  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 1762.2
##
## Number of Fisher Scoring iterations: 5
exp(confint(model0))
## Waiting for profiling to be done...
##                  2.5 %     97.5 %
## (Intercept) 0.06283227 0.08728473
## arcus1      1.25429424 2.12406175

#Alternative: Using the gtsummary library to create a table from the model above
model0 %>%
  tbl_regression(exponentiate = T)
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**OR**	**95% CI**	**p-value**
arcus			
0	—	—	
1	1.64	1.25, 2.12	<0.001

A manual calculation returns OR=1.635 which is exactly the point estimate provided by logistic regression in an unadjusted analysis, OR=1.635, 95%CI=(1.25 ; 2.12)

c) A large sample formula for the standard error of the log-OR estimate in a 2x2 table is given by: $SE(log(\hat{OR})) = \sqrt{1/a + 1/b + 1/c + 1/d)}$ where a, b, c and d are the frequencies in the 2x2 table. Compute the 95% CI for the estimate you have just computed. How does it compare with the 95% obtained from logistic regression. Hint: start by computing a 95% CI for the log-OR.

```
OR<-102*2058/(153*839) # 1.635 - same as logistic reg
SElogOR<-sqrt(1/102+1/2058+1/153+1/839)
CI1=c(log(OR)-1.96*SElogOR, log(OR)+1.96*SElogOR)
CI2=exp(CI1)
CI2
## [1] 1.256994 2.127409
```

A similar 95% CI is obtained, i.e. 95% CI=(1.26 ; 2.13). A small difference may be observed since SEs are computed using Woolf's formula in the manual calculation. The information matrix is used to compute SEs in logistic regression.

Investigation

Stata code and output

```
use wcgs.dta
tabulate agec chd69, row chi2
tabulate agec_type chd69, row chi2
scalar OR1=55*512/(1036*31)
scalar OR2=70*512/(680*31)
scalar OR3=65*512/(463*31)
scalar OR4=36*512/(206*31)
disp OR1
disp OR2
disp OR3
disp OR4
## . use wcgs.. tabulate agec chd69, row chi2
##
## +-----------------+
## | Key             |
## |-----------------|
## |   frequency     |
## | row percentage  |
## +-----------------+
##
##            |         chd69
##       agec |         0          1 |     Total
## -----------+----------------------+----------
##          0 |       512         31 |       543
##            |     94.29       5.71 |    100.00
## -----------+----------------------+----------
```

```
##           1 |     1,036         55 |     1,091 
##             |     94.96       5.04 |    100.00 
## -----------+----------------------+----------
##           2 |       680         70 |       750 
##             |     90.67       9.33 |    100.00 
## -----------+----------------------+----------
##           3 |       463         65 |       528 
##             |     87.69      12.31 |    100.00 
## -----------+----------------------+----------
##           4 |       206         36 |       242 
##             |     85.12      14.88 |    100.00 
## -----------+----------------------+----------
##       Total |     2,897        257 |     3,154 
##             |     91.85       8.15 |    100.00 
## 
##           Pearson chi2(4) =  46.6534   Pr = 0.000
## 
## . tabulate agec_type chd69, row chi2
## 
## +----------------+
## | Key            |
## |----------------|
## |   frequency    |
## | row percentage |
## +----------------+
## 
##             |         chd69
##   agec_type |         0          1 |     Total
## -----------+----------------------+----------
##       35-40 |       512         31 |       543 
##             |     94.29       5.71 |    100.00 
## -----------+----------------------+----------
##       41-45 |     1,036         55 |     1,091 
##             |     94.96       5.04 |    100.00 
## -----------+----------------------+----------
##       46-50 |       680         70 |       750 
##             |     90.67       9.33 |    100.00 
## -----------+----------------------+----------
##       51-55 |       463         65 |       528 
##             |     87.69      12.31 |    100.00 
## -----------+----------------------+----------
```

```
##       56-60 |       206          36 |       242
##             |     85.12       14.88 |    100.00
## -----------+----------------------+----------
##       Total |     2,897         257 |     3,154
##             |     91.85        8.15 |    100.00
##
##           Pearson chi2(4) =  46.6534   Pr = 0.000
##
## . scalar OR1=55*512/(1036*31)
##
## . scalar OR2=70*512/(680*31)
##
## . scalar OR3=65*512/(463*31)
##
## . scalar OR4=36*512/(206*31)
##
## . disp OR1
## .87682152
##
## . disp OR2
## 1.7001898
##
## . disp OR3
## 2.318679
##
## . disp OR4
## 2.8863138
```

There is a clear association between *chd69* and age categories (*agec*) as illustrated by an increased proportion of CHD as patients get older. The Chi2 test confirms this: Chi2=46.65, p-value = 1.801e-09. OR can be computed by hand as before. ORs are increasing with age except for the 2nd age category (1-45) OR=0.87 (non-significant different with the reference category (35-40))

2) Can you get similar results using logistic regression, how?

Yes, we can simply use *agec* (or *agec_type*) as a predictor in the logistic regression model. *agec* must be declared as a *i.agec* factor to get ORs per age category.

```
use wcgs.dta
logistic chd69 i.agec
## . use wcgs.. logistic chd69 i.agec
```

```
##
## Logistic regression                                 Number of obs =  3,154
##                                                     LR chi2(4)    =  44.95
##                                                     Prob > chi2   = 0.0000
## Log likelihood = -868.14866                         Pseudo R2     = 0.0252
##
## ------------------------------------------------------------------------------
##        chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##         agec |
##           1  |   .8768215   .2025406    -0.57   0.569     .5575563    1.378903
##           2  |    1.70019   .3800504     2.37   0.018     1.097046    2.634935
##           3  |   2.318679   .5274963     3.70   0.000     1.484545    3.621494
##           4  |   2.886314   .7462298     4.10   0.000     1.738895    4.790864
##              |
##        _cons |   .0605469   .0111989   -15.16   0.000     .0421358    .0870026
## ------------------------------------------------------------------------------
## Note: _cons estimates baseline odds.
```

3) Can you test the global effect of *agec* on *chd69*. How would you go about it?

The best way to do this in logistic regression is to use a LRT. The analysis could be adjusted or not (like here) but the principle is the same. Get the two fits and compute the LRT using the *lrtest* command. A slight difficult arises: how to we fit a model with no covariate in Stata (only the intercept); a possible way is to define a variable *one* equal to 1 and use the *noconstant* option.

```
use wcgs.dta
gen one=1
logistic chd69 one, noconstant
estimates store mod0
logistic chd69 i.agec
lrtest mod0
## . use wcgs.. gen one=1
##
## . logistic chd69 one, noconstant
##
## Logistic regression                                Number of obs =   3,154
##                                                    Wald chi2(1)  = 1385.15
## Log likelihood = -890.62187                        Prob > chi2   =  0.0000
##
## ------------------------------------------------------------------------------
```

```
##        chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          one |   .0887125    .005774   -37.22   0.000     .0780878    .1007828
## ------------------------------------------------------------------------------
##
## . estimates store mod0
##
## . logistic chd69 i.agec
##
## Logistic regression                                     Number of obs =  3,154
##                                                         LR chi2(4)    =  44.95
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -868.14866                             Pseudo R2     = 0.0252
##
## ------------------------------------------------------------------------------
##        chd69 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##         agec |
##           1  |   .8768215   .2025406    -0.57   0.569     .5575563    1.378903
##           2  |    1.70019   .3800504     2.37   0.018     1.097046    2.634935
##           3  |   2.318679   .5274963     3.70   0.000     1.484545    3.621494
##           4  |   2.886314   .7462298     4.10   0.000     1.738895    4.790864
##              |
##        _cons |   .0605469   .0111989   -15.16   0.000     .0421358    .0870026
## ------------------------------------------------------------------------------
## Note: _cons estimates baseline odds.
##
## . lrtest mod0
##
## Likelihood-ratio test
## Assumption: mod0 nested within .
##
##  LR chi2(4) =  44.95
## Prob > chi2 = 0.0000
```

The LRT value is 44.95 (df=4), $p < 0.001$ (actually p=4.08e-09) suggesting a very strong association between *agec* and *chd69*. This is consistent with the Chi2 analysis, the difference being that you can now adjust for other predictors (not done here). Note that there is no missing data in age so we don't need to worry about missingness; it's recommended to check or create a smaller dataset with only the variables of interest and no missing data; apparently *lrtest* gives you a warning if you forget!

Next we can adjust for relevant predictors, it's unlikely that such a significant association disappears after adjustment. We may also use *age* as a continous predictor in the model since the association appears fairly linear - regular increase across age categories on the log-odds scale.

R code and output

1) Association between *chd69* and *agec*

```
wcgs <- read.csv("https://www.dropbox.com/s/uc29ddv337zcxk6/wcgs.csv?dl=1")
table(wcgs$agec)
##
##    0    1    2    3    4
##  543 1091  750  528  242
table(wcgs$agec_type)
##
## 35-40 41-45 46-50 51-55 56-60
##   543  1091   750   528   242
table(wcgs$agec,wcgs$chd69)
##
##        0    1
##   0  512   31
##   1 1036   55
##   2  680   70
##   3  463   65
##   4  206   36
# row percentages
tab<-table(wcgs$agec,wcgs$chd69)
prop.table(tab, 1)
##
##              0          1
##   0 0.94290976 0.05709024
##   1 0.94958753 0.05041247
##   2 0.90666667 0.09333333
##   3 0.87689394 0.12310606
##   4 0.85123967 0.14876033
# chi2 test
chisq.test(wcgs$agec,wcgs$chd69)
##
##  Pearson's Chi-squared test
##
## data:  wcgs$agec and wcgs$chd69
## X-squared = 46.653, df = 4, p-value = 1.801e-09
# OR by hand
```

```r
OR1<-55*512/(1036*31)
OR2<-70*512/(680*31)
OR3<-65*512/(463*31)
OR4<-36*512/(206*31)
c(OR1,OR2,OR3,OR4)
## [1] 0.8768215 1.7001898 2.3186790 2.8863138


#Alternative: using gtsummary
wcgs %>%
  select(c("agec", "chd69")) %>%
  tbl_summary( by="chd69") %>%
  add_p()
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,897	**1**, N = 257	**p-value**
agec			<0.001
0	512 (18%)	31 (12%)	
1	1,036 (36%)	55 (21%)	
2	680 (23%)	70 (27%)	
3	463 (16%)	65 (25%)	
4	206 (7.1%)	36 (14%)	

There is a clear association between *chd69* and age categories (*agec*) as illustrated by an increased proportion of CHD as patients get older. The Chi2 test confirms this: Chi2=46.65, p-value = 1.801e-09. OR can be computed by hand as before. ORs are increasing with age except for the 2nd age category (1-45) OR=0.87 (non-significant different with the reference category (35-40))

2) Can you get similar results using logistic regression, how?

Yes, we can simply use *agec* (or *agec_type*) as a predictor in the logistic regression model. *agec* must be declared as a factor to get ORs per age category.

```r
out<-glm(chd69 ~ factor(agec),
         family=binomial,
         data=wcgs)
summary(out)
##
```

```
## Call:
## glm(formula = chd69 ~ factor(agec), family = binomial, data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5676  -0.4427  -0.3429  -0.3216   2.4444
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.8043     0.1850 -15.162  < 2e-16 ***
## factor(agec)1  -0.1315     0.2310  -0.569 0.569298
## factor(agec)2   0.5307     0.2235   2.374 0.017580 *
## factor(agec)3   0.8410     0.2275   3.697 0.000218 ***
## factor(agec)4   1.0600     0.2585   4.100 4.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1781.2  on 3153  degrees of freedom
## Residual deviance: 1736.3  on 3149  degrees of freedom
## AIC: 1746.3
##
## Number of Fisher Scoring iterations: 5

# OR and 95% CI
exp(out$coefficients)[2:5]
## factor(agec)1 factor(agec)2 factor(agec)3 factor(agec)4
##     0.8768215     1.7001898     2.3186790     2.8863138
exp(confint(out))[2:5,]
## Waiting for profiling to be done...
##                   2.5 %   97.5 %
## factor(agec)1 0.5613828 1.393113
## factor(agec)2 1.1072235 2.667553
## factor(agec)3 1.4971848 3.663645
## factor(agec)4 1.7402843 4.813782

#Using gtsummary for  OR and 95% CI
library(gtsummary)

out %>%
```

```
  tbl_regression(exponentiate=T)
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**OR**	**95% CI**	**p-value**
factor(agec)			
0	—	—	
1	0.88	0.56, 1.39	0.6
2	1.70	1.11, 2.67	0.018
3	2.32	1.50, 3.66	<0.001
4	2.89	1.74, 4.81	<0.001

3) Can you test the global effect of *agec* on *chd69*. How would you go about it?

The best way to do this in logistic regression is to use a LRT. The analysis could be adjusted or not (like here) but the principle is the same. Get the two fits and compute the LRT by hand or use the *anova* command.

```
reduced<-glm(chd69 ~ 1, family=binomial, data=wcgs)
full<-glm(chd69 ~ factor(agec), family=binomial, data=wcgs)
# by hand
LRT=2*(logLik(full)-logLik(reduced)) # no missing data
LRT
## 'log Lik.' 44.94642 (df=5)
pval=1-pchisq(LRT,4)
pval
## 'log Lik.' 4.079259e-09 (df=5)

# alternative1: using anova
out<-anova(reduced, full)
pval<-1-pchisq(out$Deviance[2],4)
pval
## [1] 4.079259e-09


# alternative2:  using the lrtest from lmtest library
library(lmtest)
## Loading required package: zoo
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
lrtest(reduced, full)
## Likelihood ratio test
##
## Model 1: chd69 ~ 1
## Model 2: chd69 ~ factor(agec)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   1 -890.62
## 2   5 -868.15  4 44.946  4.079e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The LRT returns a p-value, p=4.08e-09 suggesting a very strong association between *agec* and *chd69*. This is consistent with the Chi2 analysis, the difference being that you can now adjust for other predictors (not done here). Note that there is no missing data in age so we don't need to worry about missingness; it's recommended to check or create a smaller dataset with only the variables of interest and no missing data.

Next we can adjust for relevant predictors, it's unlikely that such a significant association disappears after adjustment. We may also use *age* as a continous predictor in the model since the association appears fairly linear - regular increase across age categories on the log-odds scale.