Leprosy and BCG vaccination - Solution

TRAN VAN HUNG

10/9/2019

Objectives of the analysis

Find the answer for the following question: 'What is the protection afforded by BCG against leprosy in Karonga District, Northern Malawi?'

Description of the dataset

The file chilumba.dta contains data from an unmatched case-control study of risk factors for leprosy. Between 1980 and 1984 a population of approximately 112,000 people living in Karonga District, Northern Malawi, were screened for leprosy. Individuals found to have leprosy were not followed further. The remaining population was followed until 1989. During the follow-up period 252 new cases of leprosy were identified. 1008 controls without leprosy at baseline were selected at random from the screened population. Leprosy and tuberculosis are caused by similar bacteria. Multibacillay leprosy and paucibacillary leprosy are two types of leprosy manifestations.Whether an individual had been vaccinated with BCG (vaccine against tuberculosis) was assessed by examining whether they possessed a typical BCG scar when screened. BCG was introduced into Karonga District mainly in mass vaccination campaigns in schools during the late 1970s.

Method

Univariate analysis is done to study the distribution of various variables as well as to compare them among the cases and the controls. Univariate association between BCG scar and Leprosy case is done by calculating the crude odd's ratio with 95% confidence interval (CI). Logistic regression will be performed using R software to find the association of Leprosy with BCG scar, while controlling for those variable that were found to have univariate association between the case-control status.

Data analysis

Import data with R

Describe structure of chilumba dataset

```
str(chilumba)
```

```
## 'data.frame':    1260 obs. of  8 variables:
##  $ id     : Factor w/ 1260 levels "1000475","1000573",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ caco   : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ agegrp : int  4 4 1 1 1 1 5 5 5 3 ...
##  $ sex    : int  1 0 1 0 0 0 1 0 1 1 ...
##  $ bcgscar: int  1 0 0 1 1 1 0 0 0 1 ...
##  $ school : int  4 3 2 3 2 2 2 3 2 2 ...
##  $ mbcont : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ pbcont : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "datalabel")= chr "Leprosy and BCG vaccination"
##  - attr(*, "time.stamp")= chr "29 Jan 1996 17:02"
##  - attr(*, "formats")= chr  "%9.0g" "%8.0g" "%8.0g" "%8.0g" ...
##  - attr(*, "types")= int  102 98 98 98 98 98 98 98
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
##  - attr(*, "var.labels")= chr  "Identity number" "1=case, 0=control" "Age group" "0=male, 1=female"
##  - attr(*, "version")= int 5
```

Based on the output of R, we have some information about chilumba dataset. It is a dataframe, comprises the following variables:

Table 1: Variables and code explaination

variables	explaination and code
id	identity number
caco	0=control; 1=case
agegr	age group when screened coded: 1=5-9; 2=10-14; 3=15-19; 4=20-29; 5>=30
sex	0=male; 1=female
bcgscar	presence of BCG scar: 0=no; 1=yes
school	duration of schooling: 1=none, 2=1-5 years primary, 3=6-8 years primary,4=secondary/tertiary
mbcont	household contact with multibacillary case: 0=no, 1=yes
pbcont	household contact with paucibacillary case: 0=no, 1=yes

Data summary

```
summ(chilumba)
```

```
## Leprosy and BCG vaccination
## No. of observations = 1260
##
##   Var. name obs. mean   median  s.d.    min.   max.
## 1 id        1260 630.5  630.5   363.875 1      1260
## 2 caco      1260 0.2    0       0.4     0      1
## 3 agegrp    1260 3.09   3       1.57    1      5
## 4 sex       1260 0.55   1       0.5     0      1
## 5 bcgscar   1260 0.41   0       0.49    0      1
## 6 school    1209 2.08   2       0.73    1      4
## 7 mbcont    1260 0.02   0       0.13    0      1
## 8 pbcont    1260 0.12   0       0.33    0      1
```

```
summary(chilumba)
```

```
##        id            caco         agegrp           sex
##  1000475:   1   Min.   :0.0   Min.   :1.000   Min.   :0.0000
##  1000573:   1   1st Qu.:0.0   1st Qu.:2.000   1st Qu.:0.0000
##  1001073:   1   Median :0.0   Median :3.000   Median :1.0000
##  1001552:   1   Mean   :0.2   Mean   :3.094   Mean   :0.5468
##  1002063:   1   3rd Qu.:0.0   3rd Qu.:5.000   3rd Qu.:1.0000
##  1003206:   1   Max.   :1.0   Max.   :5.000   Max.   :1.0000
```

```
##  (Other):1254
##     bcgscar           school          mbcont            pbcont
##  Min.   :0.0000   Min.   :1.000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :2.000   Median :0.00000   Median :0.0000
##  Mean   :0.4095   Mean   :2.075   Mean   :0.01667   Mean   :0.1222
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :4.000   Max.   :1.00000   Max.   :1.0000
##                   NA's   :51
```

The function summary is from the base library. It gives summary statistics of each variable. For variables such as 'agegrp', 'school', nonparametric descriptive statistics such as minimum, first quartile, median, third quartile and maximum, as well as the mean (parametric) are shown.

Missing data with variable 'school': 51 in total

Deal with missing data

There are several options for dealing with missing data, including:
i. delete the missing data (in the case of a very large dataset with very few missing values)
ii. Recoding missing values into another category
iii. replace missing values with the mean/median value of the feature in which they occur.
iv. run predictive models that impute the missing data. This should be done in conjunction with some kind of cross-validation scheme in order to avoid leakage. This can be very effective and can help with the final model. v. multiple imputation
vi. Other options such as KNN imputation

This is by far the most preferred method for imputation for the following reasons:
- Easy to use
- No/Less biases (if imputation model is correct)

Since school is a categorical data, the KNN algorithm with Hamming distance could be used. One of the most attractive features of the KNN algorithm is that it is simple to understand and easy to implement. Herein, we consider 2 approaches: one is multiple-imputation (parametric) with (mice package) and other is KNN imputation (non parametric) with DMwR package.

Multiple Imputation with MICE

```
# multiple imputation
chilum_impute <- mice(chilumba[,-1])
```

```
##
##  iter imp variable
##   1   1  school
##   1   2  school
##   1   3  school
##   1   4  school
##   1   5  school
##   2   1  school
##   2   2  school
##   2   3  school
##   2   4  school
```

```
##   2   5  school
##   3   1  school
##   3   2  school
##   3   3  school
##   3   4  school
##   3   5  school
##   4   1  school
##   4   2  school
##   4   3  school
##   4   4  school
##   4   5  school
##   5   1  school
##   5   2  school
##   5   3  school
##   5   4  school
##   5   5  school
```

```
data_imp <- complete(chilum_impute)
id <- chilumba$id
chilumba1 <- cbind(id,data_imp)
# data_imp is the dataset in which the missing values of variable school are replaced after multiple im
```

KNN Multiple-Imputation

```
attach(chilumba)
library(DMwR)
knnOutput <- knnImputation(chilumba)
knnOutput$school <- round(knnOutput$school)
#chilumba1 <- knnOutput
```

Case-control data description

```
library(DataExplorer)
attach(chilumba)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     id

## The following objects are masked from chilumba (pos = 5):
##
##     agegrp, bcgscar, caco, id, mbcont, pbcont, school, sex

## The following objects are masked from chilumba (pos = 6):
##
##     agegrp, bcgscar, caco, id, mbcont, pbcont, school, sex
```

```r
table(sex,caco)
```

```
##    caco
## sex   0   1
##   0 469 102
##   1 539 150
```

```r
table(agegrp,caco)
```

```
##       caco
## agegrp   0   1
##      1 267  32
##      2 200  30
##      3 143  19
##      4 144  47
##      5 254 124
```

```r
table(school,caco)
```

```
##       caco
## school   0   1
##      1 174  83
##      2 512 115
##      3 268  34
##      4  22   1
```

```r
table(bcgscar,caco)
```

```
##        caco
## bcgscar   0   1
##       0 534 210
##       1 474  42
```

```r
table( mbcont,caco)
```

```
##       caco
## mbcont   0   1
##      0 995 244
##      1  13   8
```

```r
table( pbcont,caco)
```

```
##       caco
## pbcont   0   1
##      0 895 211
##      1 113  41
```

```
table(school, agegrp)
```

```
##       agegrp
## school   1   2   3   4   5
##      1  47  20  14  35 141
##      2 219 115  70  63 160
##      3  33  92  68  70  39
##      4   0   3   7   8   5
```

Factors	case	control
Age group (year)		
5-9	32	276
10-14	30	200
15-19	19	143
20-29	47	144
School		
None	83	174
1-5 years primary	115	512
6-8 years primary	34	268
secondary/tertiary	1	22
Sex		
Male	102	469
Female	150	539
Presence of BCG scar		
Yes	42	474
No	210	534
Household contact with multibacillary case		
Yes	8	13
No	244	995
Household contact with paucibacillary case		
Yes	41	113
No	211	895

It should be noted that the numbers of observation for school == 4 respects to agegrp are limitted. Thus, we sould comebine groups school = 3 & school == 4 into one group. The table below describes data after combining group 3 & 4 into one group.

```
chilumba$school <- ifelse(chilumba$school == 4, 3, chilumba$school)
```

Factors	case	control
Age group (year)		
5-9	32	276
10-14	30	200
15-19	19	143
20-29	47	144
School		
None	83	174
1-5 years primary	115	512
= or > 6 years primary	35	290
Sex		

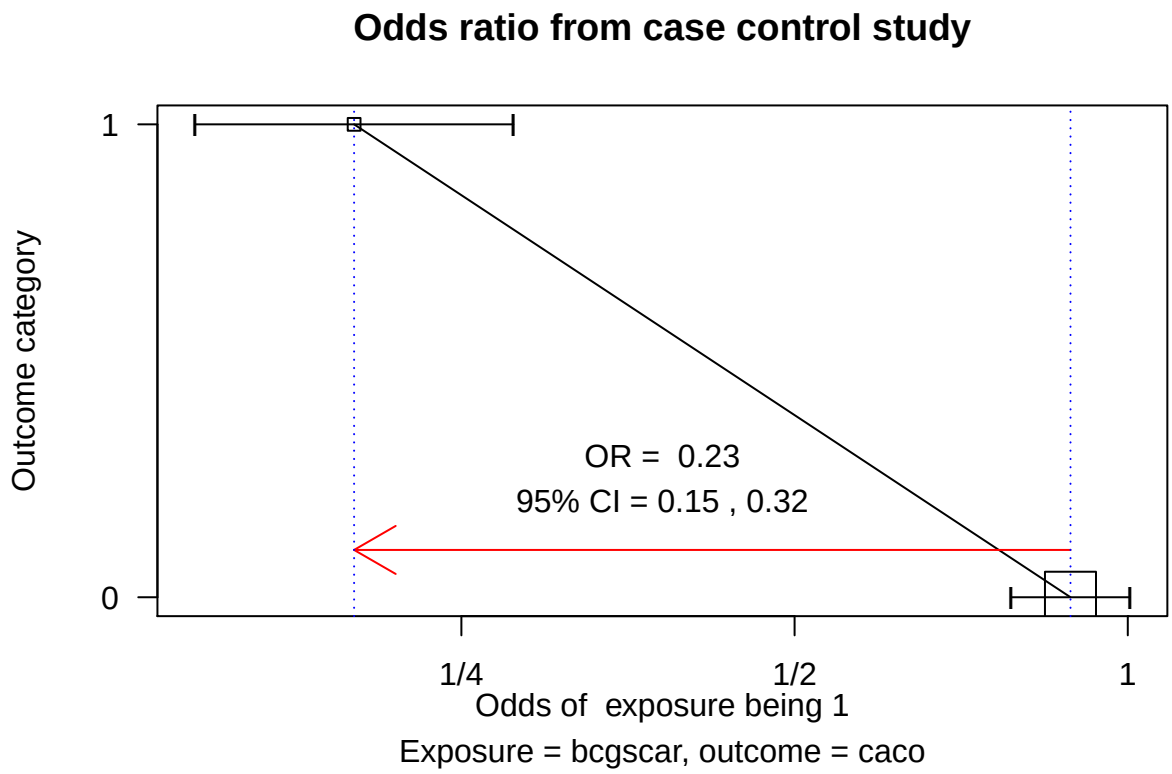
Factors	case	control
Male	102	469
Female	150	539
Presence of BCG scar		
Yes	42	474
No	210	534
Household contact with multibacillary case		
Yes	8	13
No	244	995
Household contact with paucibacillary case		
Yes	41	113
No	211	895

Testing association of BCG with leprosy

Epicalc has a function cc producing odds ratio, its 95% confidence interval, performing the chi-squared and Fisher's exact tests and drawing a graph for the explanation. It should be noted that 'cc' function uses the exact method to calculate the odds ratio.

```
library(MASS)
# table(bcgscar,caco)
# chisq.test(table(caco,bcgscar))
#

epicalc::cc(caco, bcgscar,design = "case-control")
```




```
##
##        bcgscar
## caco       0    1 Total
##   0      534  474  1008
##   1      210   42   252
##   Total  744  516  1260
##
## OR =  0.23
## Exact 95% CI =  0.15, 0.32
## Chi-squared = 76.83, 1 d.f., P value = 0
## Fisher's exact test (2-sided) P value = 0
```

The vertical lines of the resulting graph show the estimate and 95% confidence intervals of the two odds of being diseased, non-exposed on the left and exposed on the right, computed by the conventional method. The size of the box at the estimate reflects the relative sample size of each subgroup. There were more non-exposed than exposed. The non-exposed group has the estimate value slightly higher than 0.39 (true value = 210/534 = 0.393) since it real value is 210/534. The exposed group estimate is 42/474, lower than 1. The latter value is equal to 0.23 time of the former.

We observe the Pearson Chi-Squared statistic, $X_2(2) = 76.83$, corresponding to a p_value << 0.001. Therefore we have overwhelming evidence to reject the null hypothesis and thus there is strong evidence to suggest an association between leprosy case and BCG vaccination. A positive result from a chi-squared test indicates that there is some kind of relationship between two variables.

Estimating OR with a regression

```r
x = glm(caco ~ bcgscar, family=binomial(link="logit"))
require(MASS)
exp(cbind(coef(x), confint(x)))
```

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

```
##                            2.5 %    97.5 %
## (Intercept) 0.3932584 0.3345807 0.4605162
## bcgscar     0.2253165 0.1563866 0.3177336
```

OR obtained is similar.

Confounding

```r
#attach(chilumba1)
epiDisplay::mhor(caco, bcgscar,school,design = "case-control" )
```

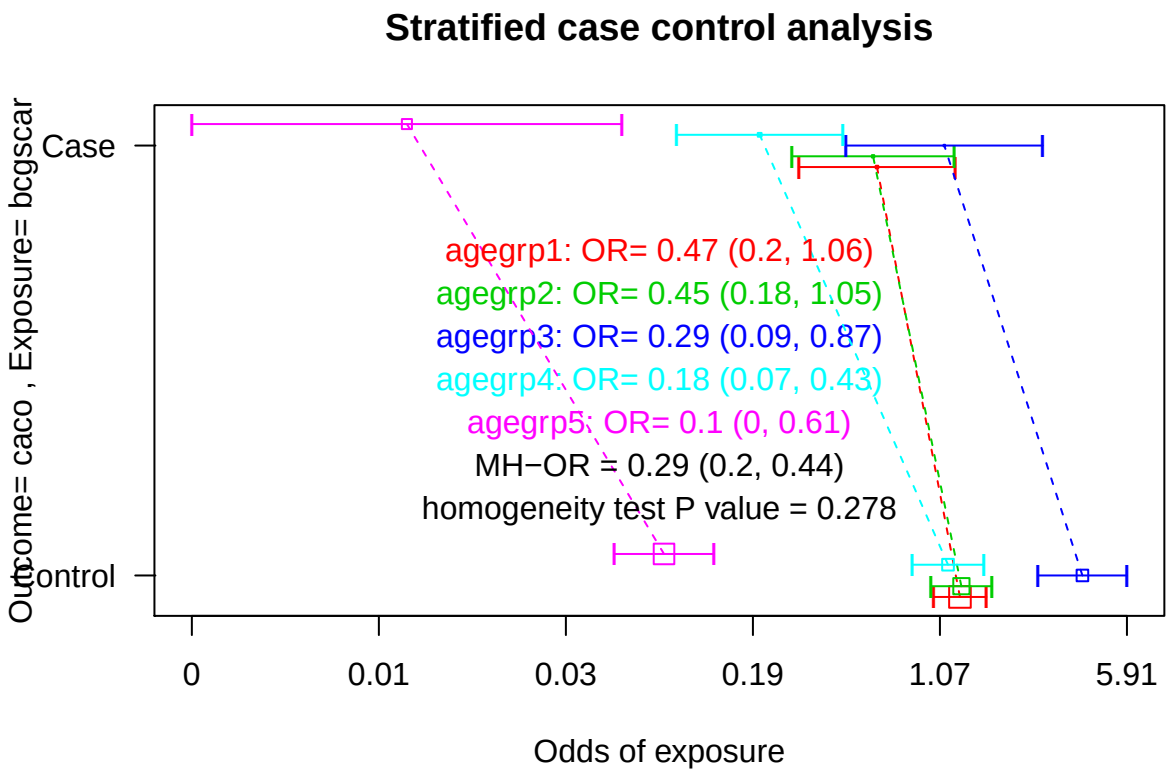
```
##
## Stratified analysis by  school
##                 OR lower lim. upper lim.  P value
## school 1     0.224     0.0747      0.561 3.62e-04
## school 2     0.331     0.1965      0.540 1.85e-06
## school 3     0.167     0.0629      0.399 7.79e-06
## school 4       Inf     0.0165        Inf 1.00e+00
## M-H combined 0.274     0.1887      0.397 6.33e-13
##
## M-H Chi2(1) = 51.74 , P value = 0
##
##  One or more cells of the stratified table == 0.
##  Homogeneity test not computable.
##
##  Graph not drawn
##
```

There are two main parts of the results. The first part concerns the odds ratio of the exposure of interest in each stratum defined by the third variable, in this case 'school' as well as the odds ratio and chi-squared statistics computed by Mantel-Haenszel's technique. The second part suggests whether the odds ratio of these strata can be combined.

The stratified analysis shows output tables for the three strata of school and corresponding three exposure lines in the graph. The odds among the control groups is higher (more on the left) compare to those of case groups. The slopes of the three lines are somewhat similar indicating minimal interaction, and this is confirmed by the P value from the homogeneity test. The MH combined odds ratio is similar to the crude odds ratio indicating rather little effect of confounding by school duration.

```r
#attach(chilumba1)
epiDisplay::mhor(caco, bcgscar,agegrp,design = "case-control" )
```

```
##
## Stratified analysis by  agegrp
##                  OR lower lim. upper lim.  P value
## agegrp 1     0.4692    0.20038      1.055 5.96e-02
## agegrp 2     0.4473    0.18210      1.049 5.01e-02
## agegrp 3     0.2854    0.09415      0.873 1.77e-02
## agegrp 4     0.1800    0.06777      0.426 8.75e-06
## agegrp 5     0.0955    0.00228      0.611 3.29e-03
## M-H combined 0.2936    0.19688      0.438 5.70e-10
##
## M-H Chi2(1) = 38.42 , P value = 0
## Homogeneity test, chi-squared 4 d.f. = 5.09 , P value = 0.278
```



The stratified analysis shows output tables for the 5 strata of age group and corresponding 5 exposure lines in the graph. The odds among the control groups is higher (more on the left) compare to those of case groups. the P value from the homogeneity test suggests a slight interaction as well as a little effect of confounding by age group. In addition, we can combine agegrp1 with age agegroup2 because they are similar lines on the figure above.

2 figures above showed no linear trend of the leprosy risk by age/school duration, it would be better to use categorical variable for age/school.

Recoding (we should test)

However, in this case, agegrp and school are categorical variables. Thus we should transform these variable from numeric (int) into categorical (factor) as the followings. In calculating an average value for a categorical

variable, a numeric value must be assigned to each category.

```
# dataset chilumba1 = corrected missing data of chilumba
# the last analysis showed
attach(chilumba1)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     id

## The following objects are masked from chilumba (pos = 3):
##
##     agegrp, bcgscar, caco, id, mbcont, pbcont, school, sex

## The following objects are masked from chilumba (pos = 6):
##
##     agegrp, bcgscar, caco, id, mbcont, pbcont, school, sex

## The following objects are masked from chilumba (pos = 7):
##
##     agegrp, bcgscar, caco, id, mbcont, pbcont, school, sex
```

```
#chilumba$agegrp <- ifelse(chilumba$agegrp==1,2,chilumba$agegrp)
chilumba1$agegrp <- ifelse(chilumba1$agegrp==1, 2,chilumba1$agegrp)

chilumba1$age <- with(chilumba1,ifelse(agegrp==1,7,ifelse(agegrp==2,12,ifelse(agegrp==3,17,
                                                                               ifelse(agegrp==4,24.5,35)))))



chilumba1$schoolnew <- with(chilumba1,ifelse(school==1,0,ifelse(school==2,3,ifelse(school==3,7,11))))




chilumba$age <- with(chilumba,ifelse(agegrp==1,7,ifelse(agegrp==2,12,ifelse(agegrp==3,17,
                                                                            ifelse(agegrp==4,24.5,35)))))



chilumba$schoolnew <- with(chilumba,ifelse(school==1,0,ifelse(school==2,3,ifelse(school==3,7,11))))
```

OR using logistic regression models

Stepwise selection of independent variables

The model may be overparametered. We let R select the model with lowest AIC. The stepwise logistic regression can be easily computed using the R function stepAIC() available in the MASS package. It performs model selection by AIC. It has an option called direction, which can have the following values: “both”, “forward”, “backward”.

```r
chilumba$agegrp <- as.factor(chilumba$agegrp)
chilumba$school <- as.factor(chilumba$school)


#md0 <- md <- glm(caco ~  bcgscar + mbcont + pbcont +sex, data = chilumba, family=binomial(link="logit"

md <- glm(caco ~ agegrp + age + schoolnew + school + bcgscar + mbcont + pbcont +sex, data = chilumba,


                  family=binomial(link="logit"))

# compare AIC of 2 models
#anova(md0,md1)
# or exactRT



modelstep <- step(md, direction = "both")
```

```
## Start:  AIC=1077.38
## caco ~ agegrp + age + schoolnew + school + bcgscar + mbcont +
##     pbcont + sex
##
##
## Step:  AIC=1077.38
## caco ~ agegrp + age + school + bcgscar + mbcont + pbcont + sex
##
##
## Step:  AIC=1077.38
## caco ~ agegrp + school + bcgscar + mbcont + pbcont + sex
##
##           Df Deviance    AIC
## - sex      1   1055.4 1075.4
## <none>         1055.4 1077.4
## - mbcont   1   1059.2 1079.2
## - pbcont   1   1061.8 1081.8
## - school   2   1070.1 1088.1
## - agegrp   4   1078.7 1092.7
## - bcgscar  1   1088.2 1108.2
##
## Step:  AIC=1075.38
## caco ~ agegrp + school + bcgscar + mbcont + pbcont
##
##           Df Deviance    AIC
## <none>         1055.4 1075.4
## - mbcont   1   1059.2 1077.2
## + sex      1   1055.4 1077.4
## - pbcont   1   1061.8 1079.8
## - school   2   1071.5 1087.5
## - agegrp   4   1078.9 1090.9
## - bcgscar  1   1088.2 1106.2
```

```r
summary(md)
```

```
##
## Call:
## glm(formula = caco ~ agegrp + age + schoolnew + school + bcgscar +
##     mbcont + pbcont + sex, family = binomial(link = "logit"),
##     data = chilumba)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6923  -0.7132  -0.4371  -0.3192   2.6665
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.406533   0.272515  -5.161 2.45e-07 ***
## agegrp2      0.368920   0.284320   1.298 0.194441
## agegrp3      0.651228   0.328152   1.985 0.047196 *
## agegrp4      1.130399   0.277577   4.072 4.65e-05 ***
## agegrp5      0.913210   0.238721   3.825 0.000131 ***
## age               NA         NA      NA       NA
## schoolnew   -0.137608   0.036544  -3.766 0.000166 ***
## school2      0.024416   0.158719   0.154 0.877742
## school3            NA         NA      NA       NA
## bcgscar     -1.150084   0.209970  -5.477 4.32e-08 ***
## mbcont       1.078364   0.524067   2.058 0.039621 *
## pbcont       0.580283   0.223365   2.598 0.009379 **
## sex         -0.006237   0.165417  -0.038 0.969921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1185.2  on 1208  degrees of freedom
## Residual deviance: 1055.4  on 1198  degrees of freedom
##   (51 observations deleted due to missingness)
## AIC: 1077.4
##
## Number of Fisher Scoring iterations: 5
```

```r
logistic.display(modelstep)
```

```
##
## Logistic regression predicting 1=case, 0=control
##
##                               crude OR(95%CI)   adj. OR(95%CI)
## Age group: ref.=1
##    2                          1.25 (0.74,2.13)  1.45 (0.83,2.52)
##    3                          1.13 (0.62,2.07)  1.92 (1.01,3.63)
##    4                          2.7 (1.63,4.46)   3.09 (1.8,5.3)
##    5                          3.85 (2.5,5.93)   2.49 (1.56,3.98)
##
## Duration of schooling: ref.=1
```

```
##    2                          0.47 (0.34,0.66)  0.68 (0.48,0.97)
##    3                          0.25 (0.16,0.39)  0.38 (0.24,0.62)
##
## 0=no, 1=yes: 1 vs 0           0.23 (0.16,0.33)  0.32 (0.21,0.48)
##
## Contact with MB case: 1 vs 0  2.29 (0.91,5.82)  2.94 (1.05,8.21)
##
## Contact with PB case: 1 vs 0  1.52 (1.02,2.27)  1.79 (1.15,2.77)
##
##                               P(Wald's test) P(LR-test)
## Age group: ref.=1                            < 0.001
##    2                          0.195
##    3                          0.046
##    4                          < 0.001
##    5                          < 0.001
##
## Duration of schooling: ref.=1                < 0.001
##    2                          0.032
##    3                          < 0.001
##
## 0=no, 1=yes: 1 vs 0           < 0.001        < 0.001
##
## Contact with MB case: 1 vs 0  0.039          0.049
##
## Contact with PB case: 1 vs 0  0.009          0.011
##
## Log-likelihood = -527.6903
## No. of observations = 1209
## AIC value = 1075.3807
```

```
# summary(md)$coef#
# OR_min <- exp(summary(md)$coef[,1] - summary(md)$coef[,2])#
# OR <- exp(summary(md)$coef[,1])
# OR_max <- exp(summary(md)$coef[,1] +  summary(md)$coef[,2])   #
# df <- data.frame(OR_min, OR, OR_max, 'p-value' = summary(md)$coef[,4])
# library(xtable)
# library(kableExtra)
# xtable(df)#
# library(sjPlot)
# kable(as.matrix(summary(md)$coef))#
# kable(as.matrix(round(df,3)))
```

with corrected missing dataset

```
library(MASS)
chilumba1$agegrp <- as.factor(chilumba1$agegrp)
chilumba1$school <- as.factor(chilumba1$school)

md2 <- glm(caco~agegrp+ bcgscar +school + mbcont + pbcont +sex+age+schoolnew , data = chilumba1, family=
modelstep2 <- step(md2, direction = "both")
```

```
## Start:  AIC=1143.19
## caco ~ agegrp + bcgscar + school + mbcont + pbcont + sex + age +
```

```
##     schoolnew
##
##
## Step:  AIC=1143.19
## caco ~ agegrp + bcgscar + school + mbcont + pbcont + sex + age
##
##
## Step:  AIC=1143.19
## caco ~ agegrp + bcgscar + school + mbcont + pbcont + sex
##
##           Df Deviance    AIC
## - sex      1   1121.4 1141.4
## <none>         1121.2 1143.2
## - mbcont   1   1126.1 1146.1
## - pbcont   1   1129.3 1149.3
## - school   3   1136.2 1152.2
## - agegrp   3   1147.7 1163.7
## - bcgscar  1   1155.1 1175.1
##
## Step:  AIC=1141.44
## caco ~ agegrp + bcgscar + school + mbcont + pbcont
##
##           Df Deviance    AIC
## <none>         1121.4 1141.4
## + sex      1   1121.2 1143.2
## - mbcont   1   1126.4 1144.4
## - pbcont   1   1129.5 1147.5
## - school   3   1136.6 1150.6
## - agegrp   3   1147.7 1161.7
## - bcgscar  1   1155.2 1173.2
```

```
logistic.display(modelstep2)
```

```
##
## Logistic regression predicting caco
##
##                   crude OR(95%CI)   adj. OR(95%CI)    P(Wald's test)
## agegrp: ref.=2
##    3              1 (0.58,1.73)     1.55 (0.87,2.76)  0.137
##    4              2.46 (1.61,3.75)  2.64 (1.68,4.14)  < 0.001
##    5              3.68 (2.61,5.17)  2.33 (1.59,3.41)  < 0.001
##
## bcgscar: 1 vs 0  0.23 (0.16,0.32)  0.32 (0.22,0.48)  < 0.001
##
## school: ref.=1
##    2              0.48 (0.35,0.66)  0.69 (0.49,0.97)  0.032
##    3              0.29 (0.19,0.44)  0.48 (0.3,0.75)   0.002
##    4              0.09 (0.01,0.7)   0.12 (0.02,0.96)  0.045
##
## mbcont: 1 vs 0   2.51 (1.03,6.12)  3.19 (1.2,8.5)    0.02
##
## pbcont: 1 vs 0   1.54 (1.04,2.27)  1.87 (1.23,2.86)  0.004
##
##                   P(LR-test)
```

```
## agegrp: ref.=2   < 0.001
##    3
##    4
##    5
##
## bcgscar: 1 vs 0  < 0.001
##
## school: ref.=1   0.002
##    2
##    3
##    4
##
## mbcont: 1 vs 0   0.026
##
## pbcont: 1 vs 0   0.005
##
## Log-likelihood = -560.7216
## No. of observations = 1260
## AIC value = 1141.4432
```

With both dataset, the variable school and agegrp should be considered as categorical variables.

Conclusion

In this analysis, the OR of presence of BCG is 0.23[CI95%: 0.16-0.33] (with/without correct missing data) suggests protective effect of BCG vaccine in prevention of leprosy. Gender differences are not significant. Also, school duration is protestive factor while housing condition is risk factor of leprosy.