EUROPEAN PROGRAMME FOR INTERVENTION EPIDEMIOLOGY TRAINING

Multivariable Analysis Module

Case Study B

An outbreak of gastroenteritis in Stegen, Germany

--Logistic regression--

March 2016
Vienna, Austria

Case study: Version 2.0 – March 2016
Authors: Alain Moren, Gilles Desvé
Modified by: Esther Kissling
Adapted to the MVA module needs: Alicia Barrasa, Ioannis Karagiannis
Based on an investigation conducted by Anja Hauri, RKI, Berlin, 1998

Copyright and License

Source:
This case study was first designed by Alain Moren and Gilles Desve, EPIET. It is based on an investigation conducted by Anja Hauri, RKI, Berlin, 1998.

Minor revisions were brought to this case study by IntoEpi 2009, 2010.

Revisions:
This is a version modified for the EPIET/EAP/EUPHEM 2015 and 2016 Multivariable Analysis modules.
Modifications:

> *2015 - Alicia Barrasa (EPIET) and Ioannis Karagiannis (PHE): The case study has been divided in two parts: the first includes descriptive, univariable and stratified analysis as pre-module homework (not shown here); the second includes logistic and binary regression.*
> *Unnecessary toponymes were removed*
> *Some sections of the help have been expanded including more explanations.*
> *2016 - Alicia Barrasa (EPIET), Ioannis Karagiannis (PHE) and Thomas Inns (PHE): The use of the `glm` command and the mathematical representation of the models have been added.*

Objectives

At the end of the case study, participants should be able to analyse data from a foodborne outbreak investigation using logistic and binomial regression, and to sort out the respective roles played by several food vehicles.

Guide to the case study

The case study is designed for use with Stata.

Nomenclature:

• `command`	Command and variable names
• **tira**	Name of dataset or file currently open

All files necessary for completing a session are placed in the corresponding session folder. There should be no need to copy files from other session folders, unless you use your own data files.

Session 1 – Logistic regression: adjusting for confounding

Remember the scenario by checking the pre-module homework.

Case study continued

Univariate and stratified analysis results suggest that tiramisù, dark and white chocolate as well as fruit salad and red jelly consumption were associated with illness (since RRs are high even among those who did not eat tiramisu). Such an association can be real (several contaminated food items, use of a single spoon to serve portions) or due to another unidentified confounding factor.

Interpretation of results should also be careful due to the small number of cases involved in this stratified analysis.

Q7. What is the next step on your plan of analysis?

Q8. Conduct a multivariable analysis using logistic regression.

Proposed steps

- Using **tiraclean.dta**, start with the `logit` command and perform a logistic regression analysis with only one exposure dichotomous variable (i.e. exposure = tiramisu, outcome = ill), interpret the results and calculate the odds

- Repeat the analysis with tiramisu as exposure, using the `logistic` command, and interpret the results

- Repeat the analysis using `tportion` as a categorical exposure variable and interpret the results

- Repeat the analysis using `tportion` as a continuous exposure variable and interpret the results

- Adding more variables to the model, discuss the meaning of the constant term for each one of them

- Write down the model for each one of the above steps

- Start again with a simple model (one independent variable) and add more variables in a step-by-step fashion

- Comparing each new nested model with the previous one (assessing the contribution of the new variable you add each time) by using the likelihood ratio test

- Assessing the fit of each model, try to identify the most parsimonious model

Help Q7

The objective of your multivariable analysis is to identify variables independently associated with the outcome and to control for confounding.

To prepare your dataset for multivariable analysis, you need to decide on the variables of interest based on your prevoious descriptive and stratified analysis and you might need to create or recode varibles (age groups, dummy variables, etc...)

Help Q8.

<u>Logistic regression using tiramisu as a dichotomous variable:</u>

Using the `logit` *command, you obtain the regression coefficient*

```
. logit ill tira, nolog

Logistic regression                               Number of obs   =        286
                                                  LR chi2(1)      =     185.04
                                                  Prob > chi2     =     0.0000
Log likelihood = -93.202968                       Pseudo R2       =     0.4982

------------------------------------------------------------------------------
         ill |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tira |   4.364143   .4436909     9.84   0.000     3.494525    5.233761
       _cons |  -3.116685   .3862464    -8.07   0.000    -3.873714   -2.359656
------------------------------------------------------------------------------
```

You can write down the above model by substituting α and β with the coefficients above.

$$\mathbf{ln}\left(\frac{\mathbf{p}}{\mathbf{1\text{-}p}}\right)=\boldsymbol{\alpha}+\boldsymbol{\beta}\mathbf{x}$$

ln(p/1-p)) is the log of the odds for the outcome
α is the log of the odds in the unexposed
β is the log of the OR for exposure x

*log odds = -3.11+(4.36 * tira)*
The OR for tiramisu is:

```
. di exp(4.364143)
78.582026
```

The `logit` *command with the* `or` *option or the* `logistic` *command (no option needed) gives you the ORs.*

```
. logistic ill tira

Logistic regression                               Number of obs   =        286
                                                  LR chi2(1)      =     185.04
                                                  Prob > chi2     =     0.0000
Log likelihood = -93.202968                       Pseudo R2       =     0.4982

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tira |   78.58201   34.86612     9.84   0.000     32.93463    187.4966
       _cons |   .0443038   .0171122    -8.07   0.000      .020781    .0944527
------------------------------------------------------------------------------
```

This model corresponds to the equation *odds = exp(α + βX) = cons*exp(βX) = cons*exp(β)*^{X}

The _cons is exp(α), which in cohort studies can be interpreted as the odds of being a case among the unexposed; in case control studies the interpretation is meaningless. Note that even if it is shown in the OR column, it is not an OR. This odds needs to be multiplied with the correct odds ratios for each exposure group to produce the odds of being a case for each exposure combination.

The OR=78.58 corresponds to exp(β) in the equation above.

Logistic regression using `tportion` *as a categorical variable:*

For categorical variables, you can create dummy variables for each level of the variable (minus 0, the reference level of that exposure).

You can directly include tiramisu portions as categorical variables when running any regression model by using the i. prefix. This will result in Stata considering variables as categorical and directly create dummy variables. The lowest value of `tportion` *is automatically set as a reference category.*

```
. logistic ill i.tportion

Logistic regression                               Number of obs     =        286
                                                  LR chi2(3)        =     193.81
                                                  Prob > chi2       =     0.0000
Log likelihood = -88.815775                       Pseudo R2         =     0.5218

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    tportion |
 One portion  |   47.29252   22.15858     8.23   0.000     18.87852    118.4723
 Two portion  |   214.4286   139.8729     8.23   0.000     59.70896    770.0622
Three portion  |   135.4286   115.9097     5.74   0.000     25.30401    724.8217
       _cons |   .0443038   .0171122    -8.07   0.000      .020781    .0944527
------------------------------------------------------------------------------
```

We can however ask Stata to change the reference level (for example use 3 instead of 0)

Use the following line of commands and check what happens:

```
char tportion[omit] 3
xi: logistic ill i.tportion
```

to change the reference level back

```
char tportion[omit]
```

Logistic regression using `tportion` *as a continuous variable:*

What would have happened if we had included `tportion` *without indicating that it is categorical?*
Try it and interpret the OR:

```
logistic ill tportion

Logistic regression                               Number of obs     =        286
                                                  LR chi2(1)        =     174.10
                                                  Prob > chi2       =     0.0000
```

```
Log likelihood = -98.668298                       Pseudo R2       =     0.4687

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    tportion |   14.21641   4.241236     8.90   0.000     7.922237    25.51128
       _cons |   .0818836   .0226111    -9.06   0.000     .0476595    .1406842
------------------------------------------------------------------------------
```

Remember that the logistic equation can be expressed as:
$odds = cons + exp(\beta X) = cons + exp(\beta)^X$

The coefficient 14.21 represents the increase in the OR with one unit increase in tportion. What would be the OR for a two-unit increase in tportion?

Adding a second variable to the model

```
. logistic ill tira beer

Logistic regression                               Number of obs   =        266
                                                  LR chi2(2)      =     172.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -88.215108                       Pseudo R2       =     0.4945

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tira |   74.02744   33.66618     9.46   0.000     30.35872    180.5103
        beer |   .4689017   .1889102    -1.88   0.060     .2128881    1.032791
       _cons |    .063345   .0255037    -6.85   0.000     .0287743    .1394503
------------------------------------------------------------------------------
```

$Odds[illness] = exp(\alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3) = exp(a)*exp(\beta_1X_1)*exp(\beta_2X_2) = \_cons*exp(\beta_1*tira)*exp(\beta_2*beer)$
Note that, in the above expression, tira and beer can get the values 0 or 1, according to whether they consumed tira or beer respectively.
`_cons` *= 0.063 is the odds of illness among the unexposed, i.e. among those who consumed neither* `tiramisu` *nor* `beer`
$exp(\beta_1) = 74.02$ *is the OR for tira adjusted by beer. The odds of illness among those who consumed tiramisu but did not drink* `beer` *is 74 times higher compared to those who consumed neither* `tiramisu` *nor* `beer`.
$exp(\beta_2) = 0.47$ *is the OR for beer adjusted by tira. The odds of illness among those who drank* `beer` *but did not consume* `tiramisu` *is almost half the odds of those who consumed neither tiramisu nor beer; however, this finding is not statistically significant.*
*The odds of illness among those who ate tiramisu and beer is 74.02*0.47 times higher than among those who consumed neither.*

Adding a third variable to the model

```
. logistic ill tira beer mousse

Logistic regression                               Number of obs   =        265
                                                  LR chi2(3)      =     175.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -86.275647                       Pseudo R2       =     0.5043

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tira |    47.7559   23.30564     7.92   0.000     18.34962    124.2873
        beer |   .5129572    .209952    -1.63   0.103     .2299781    1.144131
      mousse |   2.339514   1.000302     1.99   0.047     1.011996    5.408446
       _cons |   .0512656   .0219748    -6.93   0.000      .022129    .1187656
------------------------------------------------------------------------------
```

Try to write down the model and interpret all its coefficients.

Adding variables in a step-by-step fashion using the likelihood ratio test `lrtest` to compare different models

Variables to be included in a multivariable regression model are selected on the basis of the results of the crude analysis. Variables showing an association with the outcome and having a p-value less than 0.2 are often considered eligible. The cut-off should be chosen depending on the specific situation. Often it is between 0.25 and 0.1 but higher p-values can sometimes be justified. However, if you have any reason to believe a specific variable (exposure) should be in the model (i.e. because it might be a confounder), you should include it in the model anyway. There is no golden rule in the final inclusion of variables in a multivariable analysis model, especially in outbreak investigations.

To be able to statistically check if the inclusion of a variable improves the model significantly, the models need to have the same number of observations. If you remember for some variables we had missings, meaning that each of them have a different number of observations. You need to drop all the missings. You can do this manually:

```
drop if ill == .
drop if tira == .
…
drop if mportion == .
save tiranomissing, replace
```

or using the ado-file dropmissing that you may need to download

```
dropmissing ill tira age dmousse wmousse beer fruitsalad redjelly
tportion mportion salmon mince tomato horseradish chickenwin
roastbeef pork

save tiranomissing, replace
```

There are two possible strategies:

- *to start off with a model that includes only one independent variable and add others one by one,*
- *to start with a full model (including all variables) and, one at a time, remove variables that do not seem relevant.*

We will begin with only one independent variable.

```
. logistic ill tira

Logistic regression                               Number of obs   =        239
                                                  LR chi2(1)      =     155.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -77.253278                       Pseudo R2       =     0.5014

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tira |      74.25   34.28803     9.33   0.000     30.03438    183.5584
       _cons |   .0518519   .0200998    -7.63   0.000     .0242552     .110847
------------------------------------------------------------------------------
```

To identify whether the addition of variables contribute significantly to the model using lrtest command we need to safe the model statistics:
`estimates store m1 //`*(will store estimates in the model m1)*

Now do a second model with one additional variable (beer)

```
logistic ill tira beer
```
`estimates store m2` *// (will store estimates in the model m2)*

```
. logistic ill tira beer

Logistic regression                               Number of obs   =        239
                                                  LR chi2(2)      =     158.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -75.471113                       Pseudo R2       =     0.5129

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tira |   80.26305   38.19716     9.21   0.000     31.58123    203.9868
        beer |   .4403052   .1941901    -1.86   0.063        .1855    1.045114
       _cons |   .0687209   .0278876    -6.60   0.000     .0310215    .1522349
------------------------------------------------------------------------------

. estimates store m2
```

Now test for the difference in log likelihood

```
. lrtest m2 m1

Likelihood-ratio test                                 LR chi2(1)  =      3.56
(Assumption: m1 nested in m2)                     Prob > chi2 =    0.0590
```

If the `lrtest` *is statistically significant, this suggests that the addition of* `beer` *in the model significantly improves the likelihood of this model.*

The results of the `lrtest` *(p = 0.0590) suggest a borderline significance (at the 0.05 level) for the addition of the variable* `beer`*. Remember this might be a confounder, so this may be a sufficient reason for which you may want to keep it in the model regardless of its p-value in the likelihood ratio test.*

Then extend to other variables and store estimates in m3. Proceed similarly to extend or drop the model according to the `lrtest` *results.*

Keep or drop other variables as needed.
Take `lrtest`*, p values, magnitude of OR, and proportion of cases exposed into account in order to decide.*

Assessing the fit of each model, try to identify the most parsimonious model

You can acquire the same coefficients for the different logistic regression models using `glm` *commands. This command can be used for any generalised linear model, including logistic* *regression, as long as you specify the link function which is appropriate for the type of model (i.e. logistic regression) that you are trying to fit. Because it provides different post-estimations to the* `logistic`, `logit` *or* `regress` *commands, it is sometimes preferred. For logistic regression the line of commands will be as follows:*

```
glm ill tira, link(logit) family(binomial) eform
```

Using the `glm` *command and the post-estimation* `estat ic` *command, the value of the AIC is shown. You can compare AIC between models to decide which model is the most parsimonious; to do this, you need to save the model statistics*

```
estat ic
```

Now do a second model including beer

```
glm ill tira beer, link(logit) family(binomial)eform
estat ic
```

You can now add more variables to the model and compare the different AIC; the model with the lowest AIC value will be the most parsimonious.

Session 2 – logistic regression: including interactions

Remember that after your stratified analysis, *It seemed that consumption of* `beer` *reduced the* effect of `tiramisu` consumption on the occurrence of gastroenteritis. The RR does not significantly differ between the two strata. But, effect modification may be present.

Q9. Take interaction into account in your logistic model:

Proposed steps, using **tiranomissing.dta**

- perform a stratified analsysis using logistic regression to check for interactions
- Add an interaction term to the model.
- Does the interaction term improve the fit of the model?

Help Q9

<u>Perform a stratified analysis using logistic regression to check for interactions.</u>

Fisrt lets remember what we saw in the stratified analysis

```
ccinter ill beer, by(tira)
Number of obs =   239 , Missing =     0

 tira = Exposed
------------------------------+
        beer  Cases   Controls|
------------------------------|        Odds Ratio   0.32 [0.10-0.99]
     Exposed    25       12   |    Attrib.risk.exp   0.68 [0.01-0.90]
   UnExposed    52        8   |    Attrib.risk.pop   0.41
------------------------------+
       Total    77       20
       Exp %    32%      60%

tira = Unexposed
------------------------------+
        beer  Cases   Controls|
------------------------------|        Odds Ratio   1.00 [0.14-6.13]
     Exposed     3       58   |    Attrib.risk.exp   0.00 [-5.13-0.86]
   UnExposed     4       77   |    Attrib.risk.pop   0.00
------------------------------+
       Total     7      135
       Exp %    43%      43%

    Test of Homogeneity (M-H) : pvalue :  0.2271119

                    Crude OR for beer :  0.61 [0.33-1.09]
      MH OR for beer adjusted for tira :  0.46 [0.20-1.06]
        Adjusted/crude relative change : -24.69 %
```

You can obtain the same ORs using logistic regression:

```
logistic ill beer if tira==0
logistic ill beer if tira==1
```

Add an interaction term to the model.

We manually generate the interaction term as a new variable (`tira_beer`*) defined as the product of* `tira` *and* `beer`*. This variable equals one if* `tira` *and beer are present at the same time. Otherwise it is zero. This interaction term (variable) is inserted into the model.*

```
gen tira_beer=tira*beer

logistic ill tira beer tira_beer
```

In Stata, by typing the following, you don't need to generate the interaction manually

```
logistic ill tira##beer
Logistic regression                               Number of obs     =        239
                                                  LR chi2(3)        =     160.38
                                                  Prob > chi2       =     0.0000
Log likelihood = -74.767616                       Pseudo R2         =     0.5175

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.tira |    125.125   79.84679     7.57   0.000     35.82297    437.0454
      1.beer |   .9956897   .7799255    -0.01   0.996     .2144748    4.622444
   tira#beer |
        1 1  |   .3219004   .3021624    -1.21   0.227     .0511344    2.026422
       _cons |   .0519481   .0266401    -5.77   0.000     .0190131    .1419334
------------------------------------------------------------------------------
```

the model is:
$odds = exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3) = cons * exp(\beta_1 tira + \beta_2 beer + \beta_3 tira\_beer)$
$odds = cons * exp(\beta_1)^{tira} * exp(\beta_2)^{bee\ r} * exp(\beta_3)^{tira*beer}$
$odds = \_cons * 125.12^{tira} * 0.99^{beer} * 0.32^{tira*beer}$

The odds of illness among those who consumed tiramisu but did not drink `beer` *was 125.12 times higher compared to those who consumed neither* `tiramisu` *nor* `beer` *(exposed group: those who consumed tiramisu and did not drink beer, unexposed group=those who were not exposed to* `tiramisu` *nor* `beer`*).*

The odds of illness among those who drank beer but did not consume `tiramisu` *was almost the same compared to those who consumed neither tiramisu nor beer (OR=0.99).*

The odds of illness among those who drank `beer` *<u>and</u> consumed* `tiramisu` *was 40 times (0.32*125.13*0.99=40.1) higher compared to those who consumed neither* `tiramisu` *nor* `beer`*.*

This result can be obtained with the `lincom` *command executed after the logistic command.*

```
lincom 1.tira + 1.beer + 1.tira#1.beer
 ( 1)  [ill]tira + [ill]beer + [ill]tira#beer = 0

------------------------------------------------------------------------------
         ill | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   40.10417   24.92662     5.94   0.000     11.86118    135.5973
------------------------------------------------------------------------------
```

Cases of gastroenteritis and controls according to level of exposure to `beer` *and* `tiramisu` *consumption.*

Tira	*Beer*	*Cases*	*Controls*	*OR*
1	*1*	*25*	*12*	*40.1042*
0	*1*	*3*	*58*	*0.9957*
1	*0*	*52*	*8*	*125.125*
0	*0*	*4*	*77*	*Reference*

Does the interaction term improve the fit of the model?

First the parameters of both models have to be stored to be compared by the likelihood ratio test. Then the test is applied.

```
logistic ill tira beer
estimates store model0
logistic ill tira##beer
estimates store model1
lrtest model0 model1
```

Is the model with the interaction a better model?

Using glm commads, check that results are the same and save the model statistic

```
glm ill tira beer, link(logit) eform nolog
estat ic
glm ill tira##beer, link(logit) eform nolog
estat ic
```

Why did we calculate OR here if for the initial stratified analysis we used RR?

Optional Session 3 – Binomial regression: dealing with RRs

In the scenario presented, investigators defined a cohort study, and in univariate and stratified analysis they presented RRs.

Logistic regression provides only odds ratios. These can always be reported and they are not wrong. However, one may want to stick to risk ratios in the multivariable analysis, too. In this case, logistic regression is not appropriate.

Q10. Repeat the analysis you just did with logistic regression using now binomial regression

Proposed steps: using **tiranomissing.dta**

- Start wiht the simplest model with one exposure variable only
- Add one variable at a time and compare models
- Add the interation and interpet it

The command for binomial regresion is `binreg` and you need to indicate the option `,rr`.

Using `glm` commands you need to indicate `family(binomial)` and `link(log)` with the option `eform` to disply RR