4

Logistic Regression and Epidemiological Analyses

After having reviewed the principal measures of risk in epidemiologic and diagnostic studies, we will focus on modeling a binary variable according to numerical or binary explanatory variables based on the logistic regression model.

4.1. Measures of association in epidemiology

4.1.1. *Prognostic studies and risk measures*

In addition to `tabodds` mentioned in Chapter 2, Stata provides the command `mhodds` for case-control and cross-sectional studies. Here is an example of how to use it with the variables `low` and `smoke` analyzed in section 2.3:

```
. mhodds low smoke

Maximum likelihood estimate of the odds ratio
Comparing smoke==1 vs. smoke==0

------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------
   2.021944       4.90        0.0269         1.069897   3.821169
------------------------------------------------------------------
```

Now we will consider four age groups for the mothers and will perform a Maentel–Haenszel test to obtain an estimate of the odds ratio by controlling the age:

```
. xtile age4 = age, nq(4)
. table low smoke age4

----------------------------------------------------------------
          |             4 quantiles of age and smoke
          | ---- 1 ---    ---- 2 ---    ---- 3 ---    ---- 4 ---
      low |    0     1       0     1       0     1       0     1
----------+-----------------------------------------------------
        0 |   20    16      26    10      15     6      25    12
        1 |    8     7       8    12      10     5       3     6
----------------------------------------------------------------
```

It is possible to utilize `mhodds low smoke age4` to directly obtain the common odds ratio, but by specifying the stratification factor in an option `by()` Stata provides the estimates per stratum in addition to the common odds ratio estimate:

```
. mhodds low smoke, by(age4)

Maximum likelihood estimate of the odds ratio
Comparing smoke==1 vs. smoke==0
by age4

------------------------------------------------------------------------------
     age4 | Odds Ratio        chi2(1)         P>chi2       [95% Conf. Interval]
----------+-------------------------------------------------------------------
        1 |   1.093750           0.02         0.8856         0.32257    3.70863
        2 |   3.900000           5.50         0.0191         1.14267   13.31098
        3 |   1.250000           0.09         0.7630         0.29217    5.34783
        4 |   4.166667           3.48         0.0619         0.81731   21.24180
------------------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for age4
    ----------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ----------------------------------------------------------------
       2.138616       5.59        0.0181         1.121338   4.078767
    ----------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(3)   =    3.36
                                     Pr>chi2   =  0.3399
```

In this case, we are manipulating individual data, but this command will also work with a table of counts. For this, the weighting will be specified by reporting the

counting variable inside the `fweight=` option (see `help mhodds` for examples of how to use it).

In terms of visualization of the stratified tabulated data, it is quite possible to easily build a series of bar graphs employing the `catplot` command presented in section 2.2.1. In order to make the reading of the chart easier, it is necessary to add labels to the three variables being manipulated: `low`, `smoke` and `age4`. For the latter, we need to know the bounds of the class intervals that the `xtile` command has used. This information can be obtained from the `_pctile` command in the following manner. Note that the two extreme bounds are not included during the displaying, but based on `summarize age` we can verify the minimum and the maximum values of this variable. Note that the bounds shown below are inclusive:

```
. _pctile age, n(4)
. return list

scalars:
                  r(r1) =  19
                  r(r2) =  23
                  r(r3) =  26
```

From this, a set of labels can be created for the three variables, and the distribution of numbers corresponding to the three-dimensional array can be displayed. The option `percent` will be included with the `catplot` command when it is preferable to display proportions rather than counts (Figure 4.1):

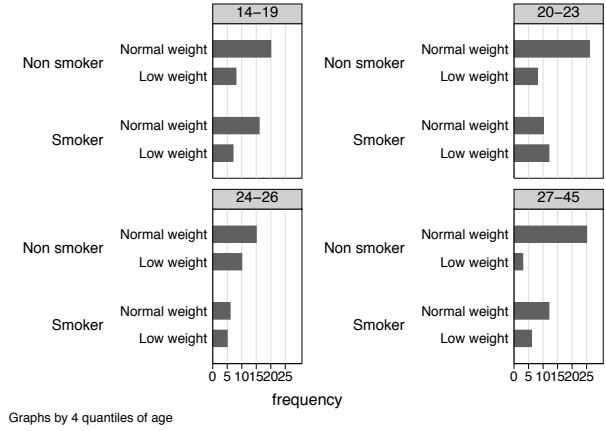

Figure 4.1. *Distribution of children with a smaller weight than the standard according to the mother's age and smoker status*

```
. label define agec 1 "14-19" 2 "20-23" 3 "24-26" 4 "27-45"
. label values age4 agec
. label define wght 0 "Normal weight" 1 "Low weight"
. label values low wght
. label define smoking 0 "Non smoker" 1 "Smoker"
. label values smoke smoking
. catplot low smoke, by(age4)
```

The `epitab` commands will provide the same result for the calculation of the odds ratio. For example, with the `cc` command for case-control studies, it would yield:

```
. cc low smoke, by(age4)

4 quantiles of an |       OR       [95% Conf. Interval]   M-H Weight
------------------+------------------------------------------------
           14-19 |    1.09375      .2719158   4.315057     2.509804 (exact)
           20-23 |        3.9       1.06682   14.50878     1.428571 (exact)
           24-26 |       1.25        .23063   6.531024     1.666667 (exact)
           27-45 |   4.166667       .713997   29.26378     .7826087 (exact)
------------------+------------------------------------------------
           Crude |   2.021944      1.029092   3.965864              (exact)
    M-H combined |   2.138616      1.130227    4.04669
-------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(3) =     3.48  Pr>chi2 = 0.3237

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      5.59
                                                Pr>chi2 =    0.0181
```

whereas, if we do not considering the `age4` variable:

```
. cc low smoke, woolf

                 | smoke                  |             Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        30          29  |         59       0.5085
        Controls |        44          86  |        130       0.3385
-----------------+------------------------+------------------------
           Total |        74         115  |        189       0.3915
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
```

```
      Odds ratio |         2.021944       |     1.08066    3.783112 (Woolf)
 Attr. frac. ex. |         .5054264       |    .0746392    .7356673 (Woolf)
 Attr. frac. pop |         .2569965       |
                 +-------------------------------------------------
                              chi2(1) =     4.92  Pr>chi2 = 0.0265
```

The response variable is always placed in the first position, followed by the exposure factor. To obtain a measure of the relative risk, `cc` will be replaced by `cs` where applicable (cohort study, or even cross-sectional studies).

The following illustrates the `low` and `smoke` variables (the risk ratio is first estimated manually using rounded percentages):

```
. tabulate low smoke, col nofreq

             |         smoke
         low | Non smoke     Smoker |     Total
-------------+----------------------+----------
Normal weight |     74.78      59.46 |     68.78
   Low weight |     25.22      40.54 |     31.22
-------------+----------------------+----------
       Total |    100.00     100.00 |    100.00

. display 40.54/25.22
1.6074544


. cs low smoke

                | smoke                |
                |   Exposed   Unexposed |      Total
----------------+----------------------+------------
          Cases |        30          29 |         59
       Noncases |        44          86 |        130
----------------+----------------------+------------
          Total |        74         115 |        189
                |                      |
           Risk |  .4054054    .2521739 |   .3121693
                |                      |
                |      Point estimate  |    [95% Conf. Interval]
                |----------------------+------------------------
 Risk difference |         .1532315     |    .0160718    .2903912
      Risk ratio |         1.607642     |    1.057812    2.443262
```

```
 Attr. frac. ex. |          .377971       |    .0546528    .5907112
 Attr. frac. pop |          .1921887       |
                 +-------------------------------------------------
                                chi2(1) =     4.92  Pr>chi2 = 0.0265
```

4.1.2. *Diagnostic studies*

With regard to the evaluation of diagnostic tests, in most cases we will start from a contingency table describing the counts associated with the cross-tabulation of two binary variables. Consider the data originating from a validation study of a new diagnostic test in 1,586 patients. Among the 744 ill patients, 670 have been identified as such by this new test.

The data are reported below. They are available in a Stata file called `diagnos.dta` and can be directly imported with the `use` command. Note that in order to avoid losing the current session, `preserve` is employed to save the data, and then, by removing all (`*`) variables, the workspace is cleansed. However, it would not be difficult for users to enter themselves the data by means of the built-in editor or the `input` command:

```
. preserve
. drop *
. use "diagnos.dta"
. list

     +------------------+
     | Test   Dis     N |
     |------------------|
  1. |    1     1   670 |
  2. |    0     1    74 |
  3. |    1     0   202 |
  4. |    0     0   640 |
     +------------------+
```

The frequency table can be reconstructed by weighting the `tabulate` command:

```
. tabulate Test Dis [fweight=N]

           |          Dis
      Test |         0          1 |     Total
-----------+----------------------+----------
         0 |       640         74 |       714
         1 |       202        670 |       872
```

```
-----------+----------------------+----------
     Total |       842        744 |     1,586
```

From this point, all the information necessary to calculate values such as sensitivity or specificity is available, as well as the positive and negative predictive values. Subsequently, we will see that it is also possible to verify the diagnostic qualities of a test based on a logistic regression model. However, there is a Stata package that automatically calculates all these quantities (use `findit diagt` and follow the installation procedures).

Here are the results that would be obtained with these data:

```
. diagt Dis Test [fw=N], chi

           |         Test
       Dis |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       670         74 |       744
    Normal |       202        640 |       842
-----------+----------------------+----------
     Total |       872        714 |     1,586

          Pearson chi2(1) = 696.4558   Pr = 0.000
True abnormal diagnosis defined as Dis = 1

                                                        [95% Confidence Interval]
-------------------------------------------------------------------------------
Prevalence                              Pr(A)       47%       44%      49.4%
-------------------------------------------------------------------------------
Sensitivity                          Pr(+|A)      90.1%     87.7%      92.1%
Specificity                          Pr(-|N)        76%       73%      78.9%
ROC area                 (Sens. + Spec.)/2         .83      .812       .848
-------------------------------------------------------------------------------
Likelihood ratio (+)     Pr(+|A)/Pr(+|N)         3.75      3.32       4.24
Likelihood ratio (-)     Pr(-|A)/Pr(-|N)         .131      .105       .163
Odds ratio                   LR(+)/LR(-)         28.7      21.5       38.2
Positive predictive value        Pr(A|+)      76.8%     73.9%      79.6%
Negative predictive value        Pr(N|-)      89.6%     87.2%      91.8%
-------------------------------------------------------------------------------
```

Care must be taken not to forget to restore the initial environment by using the appropriate command:

```
. restore
```

4.2. Logistic regression

4.2.1. *Estimation of the model parameters*

When data are available in individual format (long table where each line denotes a statistical unit for which a binary response variable and one or more explanatory variables are accessible), Stata proposes two commands for performing a single or a multiple logistic regression: `logit` and `logistic`. They mainly differ in the format of the results they display: by default, `logistic` returns odds ratios, whereas `logit` returns the regression coefficients on the log-odds scale.

Note that the `probit` command allows for the estimation of the parameters of a logistic regression model by considering a link function such as probit rather than logit.

The following gives the result of a logistic regression considering the babies' birth weight indicator (`low`) as a response variable and the mothers's weight (`lwt`) as an explanatory variable:

```
. logistic low lwt

Logistic regression                               Number of obs   =        189
                                                  LR chi2(1)      =       5.98
                                                  Prob > chi2     =     0.0145
Log likelihood = -114.34533                       Pseudo R2       =     0.0255

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |   .9695452   .0131597    -2.28   0.023     .9440927     .995684
       _cons |   2.713702   2.131045     1.27   0.204     .5822659    12.64745
------------------------------------------------------------------------------
```

As in the case of the linear regression, Stata provides the values of the model parameters (here, in terms of the odds ratio associated with the unit variation of the numeric explanatory variable) and their confidence intervals, accompanied by usual significance tests (for the coefficients and the model, `LR chi2(1)`). The value of pseudo R^2 reported by Stata corresponds to the McFadden coefficient.

Additional goodness of fit indices of the model will be obtained by making use of the `fitstat` command discussed in Chapter 4:

```
. fitstat

Measures of Fit for logistic of low
```

```
Log-Lik Intercept Only:       -117.336   Log-Lik Full Model:          -114.345
D(187):                        228.691   LR(1):                          5.981
                                         Prob > LR:                      0.014
McFadden's R2:                   0.025   McFadden's Adj R2:              0.008
ML (Cox-Snell) R2:               0.031   Cragg-Uhler(Nagelkerke) R2:     0.044
McKelvey & Zavoina's R2:         0.053   Efron's R2:                     0.032
Variance of y*:                  3.475   Variance of error:              3.290
Count R2:                        0.688   Adj Count R2:                   0.000
AIC:                             1.231   AIC*n:                        232.691
BIC:                          -751.516   BIC':                          -0.740
BIC used by Stata:             239.174   AIC used by Stata:            232.691
```

In the case where the explanatory variable is binary, the modeling principle remains the same. The following illustrates the babies' weight and the presence of intrauterine pain for the mother, this time with the `logit` command:

```
. logit low ui, nolog

Logistic regression                             Number of obs   =        189
                                                LR chi2(1)      =       5.08
                                                Prob > chi2     =     0.0243
Log likelihood = -114.79795                     Pseudo R2       =     0.0216

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ui |   .9469277   .4167734     2.27   0.023     .1300669    1.763789
       _cons |  -.9469277   .1756215    -5.39   0.000     -1.29114   -.6027159
------------------------------------------------------------------------------
```

Note that the option `or` can be added to obtain the odds ratio, which can be also found by taking the exponential of the regression coefficient stored in `_b[ui]`:

```
. display exp(_b[ui])
2.5777778
```

4.2.2. *Logistic regression and diagnostic studies*

In the case of a diagnostic study, or more generally when there is interest in a classification approach rather than a regression one, the following command yields a contingency table summarizing the individuals correctly or incorrectly classified as positive and negative with respect to the response variable, considering a threshold of 0.5 for the detection or events allocation probability:

```
. estat classification

Logistic model for low

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        14             14  |         28
     -     |        45            116  |        161
-----------+--------------------------+-----------
   Total   |        59            130  |        189

Classified + if predicted Pr(D) >= .5
True D defined as low != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   23.73%
Specificity                     Pr( -|~D)   89.23%
Positive predictive value       Pr( D| +)   50.00%
Negative predictive value       Pr(~D| -)   72.05%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   10.77%
False - rate for true D         Pr( -| D)   76.27%
False + rate for classified +   Pr(~D| +)   50.00%
False - rate for classified -   Pr( D| -)   27.95%
--------------------------------------------------
Correctly classified                        68.78%
--------------------------------------------------
```

The `lroc` command allows for displaying the Receiver Operating Characteristic (ROC) curve of the last estimated model as well as the value of the area under the curve.

4.2.3. *Point and interval prediction*

The `predict` command is used to calculate the values fitted for a given model or to estimate probability or log-odds values for new observations: this is a postestimation command, and it can therefore be utilized after having built a regression model with `logit` or `logistic`. The options make it possible to define the type of prediction that we are interested in: when the aim consists of predicting probabilities, the option `p` will be employed; otherwise, the option `xb` will provide predictions on the link scale:

```
. quietly: logit low lwt
. predict pr, p
```

The calculation of the 95% confidence intervals for fitted values poses no real difficulty. Manually, we can proceed as follows (on the log-odds scale): as a first step, we generate the linear predictions (`lo`), the prediction error (`lose`) and the bounds of the associated confidence intervals (`lolci` and `louci`). A scatterplot of the predicted values (log odds) according to the mother's weight is then displayed, upon which are superimposed lines delimiting the confidence intervals. It should be observed that it is necessary to order the coordinates of the points in the latter case (Figure 4.2):

```
. predict lo, xb
. predict lose, stdp
. gen lolci = lo - 1.96*lose
. gen louci = lo + 1.96*lose
. twoway (scatter lo lwt, sort connect(l)) (line lolci louci lwt,
            sort pstyle(p3 p3)), ///
  xlabel(35(15)120)
```

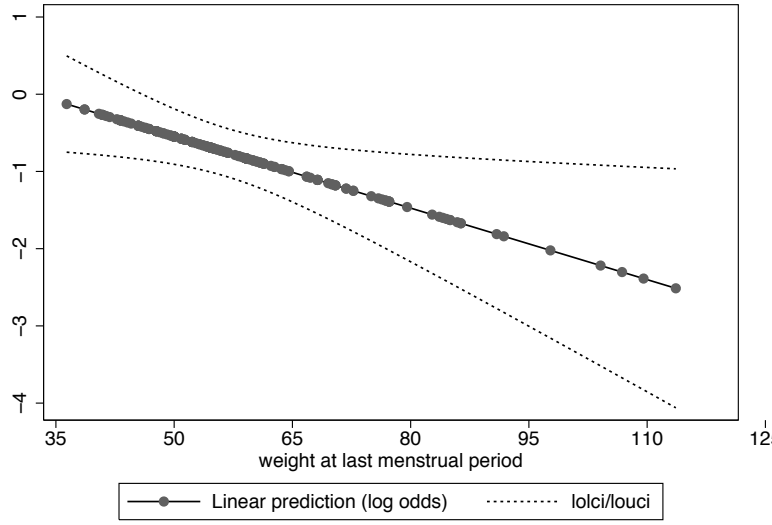

Figure 4.2. *Predicted values for the logistic regression model*

Now consider a model including, in addition to the variable `lwt`, the mothers' ethnicity (`race`). The `margins` command provides very powerful tools to calculate predictions with confidence intervals or marginal effects. Here is a relatively simplified example of how to use it for a model including these two explanatory variables, `lwt` and `race` (Figure 4.3):

```
. logit low lwt i.race
```

```
Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -111.73378
Iteration 2:   log likelihood = -111.62959
Iteration 3:   log likelihood = -111.62954
Iteration 4:   log likelihood = -111.62954

Logistic regression                               Number of obs   =        189
                                                  LR chi2(3)      =      11.41
                                                  Prob > chi2     =     0.0097
Log likelihood = -111.62954                       Pseudo R2       =     0.0486

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0334908   .0141666    -2.36   0.018    -.0612568   -.0057248
             |
        race |
          2  |   1.081066   .4880522     2.22   0.027     .1245015    2.037631
          3  |   .4806032   .3566737     1.35   0.178    -.2184644    1.179671
             |
       _cons |   .8057537   .8451667     0.95   0.340    -.8507426     2.46225
------------------------------------------------------------------------------

. quietly: margins, at(lwt=(40(10)110)) over(race)
. marginsplot

  Variables that uniquely identify margins: lwt race
```

4.2.4. *Case of grouped data*

In the case where the data are grouped, that is when there is only a table available, which summarizes the counts observed for the cross-tabulation of the levels of two categorical variables, we will use the `blogit` command instead of `logit`. It is then essential to clarify, in order, the positive events, the total number of events and the associated explanatory variables. Hereafter follows an application example with the same data (`low` and `ui`).

This time, instead of using an external file or building the data table with `input`, we will directly operate on the individual data available in the workspace. All of these manipulations will be supported by a pair of instructions `preserve`/`restore` so as to be able to return to the initial data at the end of the example.

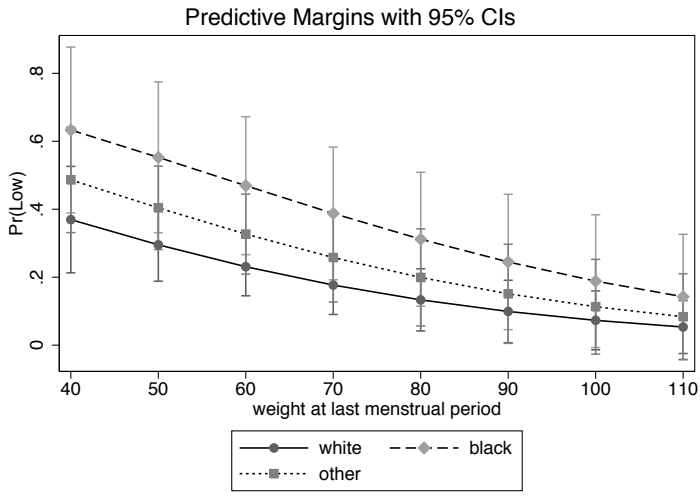

Figure 4.3. *Values predicted with the command* `margins`

Initially, a table is built summarizing the occurrence frequency of each pair of modality for the `low` and `ui` variables by making use of the `contract` command. The `freq(n)` option simply allows the variable used for counting statistics to be renamed. We will then need the total counts associated with each modality of the `ui` variable, which will allow us to obtain the sum of the positive events for the `low` variable (contained in `n`) and the sum of the positive and negative events associated (we call `tot`) when `low=1` ("Low weight"). This can be achieved with the instruction `egen`, grouped on the modalities of `ui`:

```
. preserve
. contract low ui, freq(n)
. egen tot = sum(n), by(ui)
. list

     +---------------------------------+
     |           low    ui     n   tot |
     |---------------------------------|
  1. | Normal weight    No   116   161 |
  2. | Normal weight   Yes    14    28 |
  3. |    Low weight    No    45   161 |
  4. |    Low weight   Yes    14    28 |
     +---------------------------------+
```

It is possible to verify that the total counts are correctly identical for each of the modalities of `ui`, and that the data of interest are the last two rows of this table. This enables for each level of the `ui` variable that the number of positive events and the number of total events be obtained. The regression model is then formulated as following:

```
. blogit n tot ui if low == 1, or

Logistic regression for grouped data              Number of obs   =        189
                                                  LR chi2(1)      =       5.08
                                                  Prob > chi2     =     0.0243
Log likelihood = -114.79795                       Pseudo R2       =     0.0216

------------------------------------------------------------------------------
    _outcome | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ui |   2.577778   1.074349     2.27   0.023     1.138905      5.8345
       _cons |    .387931    .068129    -5.39   0.000     .2749573    .5473232
------------------------------------------------------------------------------
```

4.3. Key points

– The `regress` command is employed in the context of the linear regression, whereas for the logistic regression, `logit` (or `logistic`) will be used. In both cases, there is a same group of postestimation commands making it possible to obtain information about the goodness of fit of the model (`fitstat`). Moreover, they can also calculate fitted values or predict new observations (`predict`).

– Most measures of risk encountered in epidemiology are accessible with the `epitab` (`cc`, `cs`), or with specific commands (`tabodds`, `mhodds`).

4.4. Further reading

As in the case of the linear regression, the book by Vittinghoff *et al.* [VIT 05] is recommended to deepen the modeling key stages of binary data. For more specific operations on categorical data with Stata, the textbook [HAR 12] is recommended.

4.5. Applications

1) We study the effect of a prophylactic therapy of a macrolide with low doses (Therapy A) on the infectious episodes in patients suffering from cystic fibrosis in a placebo-controlled multicenter randomized trial (B). The results are summarized in Table 4.1:

	Infection		
	No	Yes	Total
Therapy (A)	157	52	209
Placebo (B)	119	103	222
Total	276	155	431

Table 4.1. *Prophylaxis and cystic fibrosis therapy*

– based on a χ^2 test, how can the following question be addressed: can the treatment prevent that infectious episodes occur (considering $\alpha = 0.05$)? Verify that the expected counts are effectively larger than 5;

– can the same conclusion be achieved considering the confidence interval of the odds ratio associated with the treatment effect?

– we aim to verify whether there is a disparity from the point of view of the percentages of infectious episodes according to the center. The data per center are indicated in the table hereafter. Conclude based on a χ^2 test;

	Infection		
	No	Yes	Total
Therapy (A)	51	8	59
Placebo (B)	47	19	66
Total	98	27	125

Center 1

	Infection		
	No	Yes	Total
Therapy (A)	91	35	126
Placebo (B)	61	71	132
Total	152	106	258

Center 2

	Infection		
	No	Yes	Total
Therapy (A)	15	9	24
Placebo (B)	11	13	24
Total	26	22	48

Center 3

– based on the previous table, we aim to verify if the treatment effect is independent of the center or not. We are suggesting to carry out a comparison test between the two treatments controlling for any center effect (Mantel–Haenszel test). Indicate the result of the test as well as the value of the common odds ratio.

First of all, it is necessary to construct the counts table given in the text. A manual input method will serve as a model. On the other hand, we will directly input the labels of the variables and not those of the numeric codes. To do this, it is necessary to indicate to Stata what is the format of these labels by means of the instruction `str` which is appended with the number of characters that we want to use for the encoding of the modalities of the variables:

```
. clear all
. input str1 treatment str3 infection N

  treatment~t  infection  N
  1. "A" "No" 157
  2. "B" "No" 119
  3. "A" "Yes" 52
```

```
  4. "B" "Yes" 103
  5. end
. list
```

Here follows the statement table, with the associated χ^2 test:

```
. tabulate treatment infection [fweight=N], chi
```

If a more accurate value is desired for the degree of significance of the test, the information returned (in a non-invasive manner) by the previous command can be used:

```
. return list
```

Hence the value of the degree of significance p:

```
. display %10.9f r(p)
```

The above command instructs Stata to display the result in the form of a number with nine decimal places.

Concerning the theoretical counts, the same command as before is employed exactly but adding the option `expected`:

```
. tabulate treatment infection [fweight=N], chi expected nofreq
```

Note that the observed counts are not being displayed by means of the option `nofreq`.

In order to calculate the value of the odds ratio, the command `cc` will be entered. However, the epitab commands require that the variables be encoded in binary format (0 = non-exposed/non-sick, 1 = exposed/sick). The data entry achieved in the previous step being not compatible with this format, it is necessary to recode the data, for example by utilizing the `input` command:

```
. input treatment infection N

  treatm~t  infection N
  1. 1 0 157
  2. 0 0 119
  3. 1 1 52
  4. 0 1 103
  5. end
. label define tx 0 "Placebo" 1 "Treatment"
. label values treatment tx
. label define yesno 0 "No" 1 "Yes"
. label values infection yesno
```

Then, it is possible to perform the estimate of the odds ratio:

```
. cc infection treatment [fweight=N], woolf exact
```

It should be noted that this time the `exact` option has been included in order to obtain a Fisher's test instead of the approximation by the χ^2 distribution for the statistical test of the null hypothesis.

When examining the data per center, it is necessary to rebuild the counts tables, only considering the column margins of the count tables given in the text statement. Inputting data with `input` raises no particular difficulties:

```
. input infection centre N

  infection  centre  N
  1. 0 1 98
  2. 1 1 27
  3. 0 2 152
  4. 1 2 106
  5. 0 3 26
  6. 1 3 22
  7. end
. tabulate infection centre [fweight=N], chi
```

Once more we adopt the quick format of aggregated data entry: two variables (infectious episodes yes/no, center number) and the counts associated with the cross-tabulation of each of the levels of these variables. The `fweight` option then allows for applying a χ^2 test with the `tabulate` command weighting the two-entry table by the counts `N`.

To achieve a Mantel–Haenszel test, it is necessary to reconsider all of the data (three 2×2) tables and to make use of the `cc` command specifying the stratification factor with the option `by`. What follows is one way to proceed, taking into account three variables: `tx` (treatment A (1) or B (0)), `inf` (infection no (0)/yes (1)) and `cen` (center, 1 to 3). Care should be taken to correctly encode the classes of the two classification variables into 0 and 1:

```
. input tx inf cen N

  tx  inf  cen  N
  1. 1 0 1 51
  2. 1 1 1 8
  3. 0 0 1 47
  4. 0 1 1 19
--%<----
 13. end
```

It is possible to verify the structure of the data by employing the `table` command, by exactly following the same principle for the options (classification variables, weighting factor and stratification variable). In parallel, we will take the opportunity to add more informative labels to the modalities of the classification variables:

```
. label define txlab 0 "A" 1 "B"
. label define inflab 0 "No" 1 "Yes"
. label values tx txlab
. label values inf inflab
. table tx inf [fw=N], by(cen)
```

With respect to the Mantel–Haenszel test, we obtain the following results:

```
. cc tx inf [freq=N], by(cen)
```

2) Table 4.2 summarizes the proportion of myocardial infarctions observed in men aged 40–59 and for whom the blood pressure and the rate of cholesterol have been measured, considered in the form of ordered classes [EVE 01].

	Cholesterol (mg/100 ml)						
TA	< 200	$200 - 209$	$210 - 219$	$220 - 244$	$245 - 259$	$260 - 284$	> 284
< 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
$117 - 126$	0/66	2/27	1/25	8/69	0/24	5/22	1/19
$127 - 136$	2/59	0/34	2/21	2/83	0/33	2/26	4/28
$137 - 146$	1/65	0/19	0/26	6/81	3/23	2/34	4/23
$147 - 156$	2/37	0/16	0/6	3/29	2/19	4/16	1/16
$157 - 166$	1/13	0/10	0/11	1/15	0/11	2/13	4/12
$167 - 186$	3/21	0/5	0/11	2/27	2/5	6/16	3/14
> 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Table 4.2. *Infarction and blood pressure*

The data are available in the `hdis.dat` file in the form of a table comprising four columns that indicate, respectively, the blood pressure (eight categories, rated 1–8), the cholesterol rate (seven categories, rated 1–7), the number of myocardial infarction and the total number of individuals. We are interested in the association between blood pressure and the likelihood of suffering a myocardial infarction:

– calculate the proportions of myocardial infarction for each level of blood pressure and represent them in a table and graphical form;

– based on a logistic regression model, determine whether there is a significant association at $\alpha = 0.05$ between blood pressure, assumed as a quantitative variable considering the class centers, and the likelihood of having a heart attack;

– express in logit units, the probabilities of infarction predicted by the model for each of the blood pressures;

– on the same graph, display the empirical proportions and the logistic regression curve according to the blood pressure values (class centers).

Since the data are presented in the form of a text file in which the fields are separated by spaces, we will enter the `import delimited` command. Caution must be taken to specify that the field separator is composed of one or more spaces, and that the name of the variables appear on the first line of the file:

```
. import delimited "hdis.dat", delimiter(space, collapse) varnames(1)
```

After having verified that the data have been successfully imported, labels can be associated with the numerical codes of the `bpress` variable as follows:

```
. label define bpress 1 "<117" 2 "117-126" 3 "127-136" 4 "137-146" 5 "147-156" 6
            "157-166" 7 "167-186" 8 ">186"
. label values bpress bpress
```

The proportion of events is simply calculated by dividing the `hdis` variable by the `total` variable. Since in this exercise, there is only interest in the systolic pressure, we can immediately aggregate the data as a whole by means of `collapse`. At the same time, it can also be verified that the above calculations are accurate:

```
. collapse (sum) hdis (sum) total, by(bpress)
. generate prop = hdis/total
. list
```

To graphically represent these proportions according to the values of `bpress`, it is possible to make use of a bar or of a dot plot. For example:

```
. graph dot (asis) prop, over(bpress)
```

To convert the `bpress` variable into a numeric variable, we can proceed in the following manner:

```
. recode bpress (1 = 111.5) (2 = 121.5) (3 = 131.5) (4 = 141.5) (5 = 151.5)
            (6 = 161.5) (7 = 171.5) (8 = 181.5)
```

The logistic regression model for grouped data is then written as:

```
. blogit hdis total bpress
```

The `predict` command will then enable us to generate the values predicted by such a model for each observed predictor value (`bpress`). It will be possible to use a dot plot similar to the one utilized above to characterize the empirical proportions in order to represent the probability of myocardial infarction based on blood pressure.

3) A case-control investigation has focused on the relationship between alcohol and tobacco consumption and esophageal cancer in humans (study “Ille and Villaine”).

The group of cases consisted of 200 patients suffering from esophagus cancer and diagnosed between January 1972 and April 1974. In total, 775 male controls have been selected from the electoral lists. Table 4.3 shows the distribution of all the subjects according to their daily consumption of alcohol, bearing in mind that a consumption greater than 80 g is considered to be a risk factor [BRE 80]. The data are available in the `cc_oesophage.csv` file:

	Alcohol intake (g/day)		Total
	≥ 80	< 80	
Cases	96	104	200
Controls	109	666	775
Total	205	770	975

Table 4.3. *Infarction and blood pressure*

– what is the value of the odds ratio and its 95% confidence interval (Woolf method)? Does it provide a good estimate of the relative risk?

– is the proportion of consumers at risk the same among cases and among controls (consider $\alpha = 0.05$)?

– build the logistic regression model that enables testing the association between alcohol consumption and the status of individuals. Is the regression coefficient significant?

– recover the value of the observed odds ratio, in the preceding model, and its confidence interval based on the results of the regression analysis.

The data have been saved in compact format (three columns indicating the presence or not of cancer, the level of alcohol intake and the associated counts):

```
. insheet using "cc_oesophage.csv", clear
. label define yesno 0 "No" 1 "Yes"
. label values cancer yesno
. label define dose 0 "< 80g" 1 ">= 80g"
. label values alcohol dose
. list
```

The previous commands allow loading the data file and that more informative names be associated to the variables modalities (`cancer` and `alcohol`). To obtain the total number of patients, the following command can be utilized:

```
. egen ntot = sum(patients)
. display ntot
```

Yet once, this command is not optimal and we will prefer solutions based on the use of `scalar` and that operate on the results returned by commands such as `summarize` or `tabulate`. The `ntot` variable can be suppressed by typing:

```
. drop ntot
```

The proportion of individuals at risk, which is having a daily consumption of alcohol ≥ 80 g is obtained from a simple count table cross-tabulating the variables `cancer` and `alcohol` (it is necessary to indicate how the cells have to be filled in by adding the option `weight`), and requesting the row profiles, that is to say, the relative frequencies per row:

```
. tabulate cancer alcohol [fweight=patients], row
```

The calculation of the odds ratio can be done by means of one of the so-called "immediate" Stata command, or using `cc`, bearing in mind in this case that the `patients` column must be taken into account, as previously:

```
. cc cancer alcohol [fweight=patients], woolf
```

If the counts table, that is the distribution of the 975 subjects over the four table cells cross-tabulating the exposure to alcohol and the case-control status, is available it is also possible to make use of the case-control odds ratio calculator accessible by the Stata menus.

To test the hypothesis that the proportion of persons with a daily alcohol intake ≥ 80 g is identical in the cases (p_1) and the controls (p_0), the following method can be adopted:

```
. prtesti 96 0.4800 109 0.1406
```

Another way to test this hypothesis consists of noting that the previous hypothesis, $H_0 : \pi_0 = \pi_1$, is true only if the odds ratio is equal to 1, hence the idea of directly performing the χ^2 test for the odds ratio given by the command `cc` (here, $\chi^2 = 110.26$, $P < 0.001$).

The logistic regression model, similarly to other regression models in Stata, is formulated as follows:

```
. logistic cancer alcohol [freq=patients]
```

By default, Stata presents the results (model coefficients) in the form of odds ratio, with their associated confidence intervals. If it is desirable to directly obtain the regression coefficients (on the log-odds scale), we must have to enter the `logit` command after using `logistic`. It would also be possible to directly employ a command of the following type:

```
. logit cancer alcohol [freq=patients]
```