University of Pretoria
Data analysis for evaluation studies
Examples in STATA version 11

List of data sets

```
b1.dta   (To be created by students in class)
fp1.xls  (To be provided to students)
fp1.txt   (To be converted from Excel to Text file)
fp1.dta (To be imported into STATA by students)
fp2.dta (To be created by students)
```

How to set the main menu of STATA to default factory settings standards

Go to the main menu of STATA 11

Click on **Edit**

Click on **Preferences**, **Manage preferences**, **Load preferences**, **Factory settings**

At this stage, the main menu of STATA 11 should be set according to default factory standards

How to create a STATA log file

A STATA log file saves all STATA commands + results of data analysis.

To open a STATA log file,
Enter the main menu of STATA
Click on file, log, begin
Save in? **c:\bos**
File name? **c:bos\note1.smcl (All STATA log files have the extension .smcl)**
At this stage, you should see **LOG ON** at the bottom right corner of your screen
Keep working in STATA

To close your STATA log file,

Click on file, log, close

To translate your STATA log file into an MS Word document,

Click on file, log, translate
Input file? Click on Browse, **c:\bos\note1.smcl**
Output file? Click on Browse, **c:\bos\note1.doc** (MS Word format)
Click on translate, Read translation report

To open your translated file in MS Word,

Go to MS Word
Open file **c:\bos\note1.doc** as you would open any other MS Word document
Edit your MS Word file
Save your edited MS Word file as **c:\bos\note1.doc**

How to create a STATA dataset containing the following variables of study

Age in years (25.35 years)
Weight in Kg (56.25 Kg)
Sex (male, female)
Social support index (1, 2, 3)

Age	Weight	Sex	Social
25.35	56.25	Male	1
32.05	65.52	Female	2
44.55	66.35	Female	3
38.95	65.75	Male	2
28.69	43.27	Male	1

Type the following STATA command, and press Enter

```
. clear
```

Click on the **data editor (Edit)** icon

Enter values of the variable **age** in the **first** column of the data editor
Enter values of the variable **weight** in the **second** column of the data editor
Enter values of the variable **sex** in the **third** column of the data editor
Enter values of the variable **social** in the **fourth** column of the data editor

Click on the **Variables Manager** icon
Click on **Var1** in the Variables Manager dialog box
Overwrite **var1** by **age** in the Variables Manager dialog box
Assign a label for age as: **age in years** in the same dialog box
Click on **Apply** in the same dialog box

Click on **Var2** in the Variables Manager dialog box
Overwrite **var2** by **weight** in the Variables Manager dialog box
Assign a label for weight as: **Weight in Kg** in the same dialog box
Click on **Apply** in the same dialog box

Click on **Var3** in the Variables Manager dialog box
Overwrite **var3** by sex in the Variables Manager dialog box
Assign a label for sex as: **Gender of respondent** in the same dialog box
Click on **Apply** in the same dialog box

Click on **Var4** in the Variables Manager dialog box
Overwrite **var4** by social in the Variables Manager dialog box
Assign a label for social as: Social support index in the same dialog box
Click on **Apply** in the same dialog box

To save your data file in STATA format in the directory c:\bos using the file name **b1.dta,**

Close the Variables Manager window by clicking on inner X
Close the Data Editor (Edit) window by clicking on outer X
Click on File
Click on Save as
Save in? **c:\bos\b1.dta**, and press enter

To see the contents of the data file **c:\bos\b1.dta**,

. use c:\bos\b1.dta, clear

. list age weight sex social (shows observations in all 5 rows)

. list age weight sex social in 2/4 (shows observations in rows 2, 3 and 4 only)

How to recode in STATA

```
. use c:\bos\b1.dta, clear

. recode social  1=3  2=2  3=1

. save c:\bos\b1.dta, replace
```

Commonly used procedures in STATA for moderate level data analysis in evaluation studies

- **Importing data sets from Excel into STATA**
- **Recoding, labelling, generating new variables**
- **Summary statistics for continuous variables**
- **Frequency tables for discrete variables**
- **Pearson's chi-square tests of association**
- **Calculation of crude odds and risk ratios**
- **Binary logistic regression analysis**

How to import an Excel data set into STATA

1. Open up the data set fp1.xls in Excel, and see the content of the data set.
2. In Excel, save the data set fp1.xls as a **text tab-delimited (*.txt)** data set. Use the file name fp1.txt.
3. Go to STATA, and import your data set using the insheet command of STATA.

Click on file
Click on Import data
Click on ASCII data created by a spreadsheet
Click on Browse, and indicate the location of the file fp1.txt (Example: **f:\fp1.txt** if your file is on a flash disk F)

Click on OK

Alternatively, type the following STATA command, and press Enter

. insheet using f:\fp1.txt

(10 vars, 1333 obs)

See the contents of your imported data set in STATA by clicking on the data editor.

Assign labels to each variable of study in STATA.

Example: To assign a label to the variable age, type the following STATA command, and press Enter

label var age "Age of respondent in years"

Save your data set in STATA format

Click on file
Click on save as

Example: **save f:\fp1.dta**

Analysis of the data set fp1.dta in STATA

A survey is conducted based on a random sample of n = 474 women who took part in the study. The objective of study is to identify factors that influence the use of modern family planning methods. Raw data was gathered by using a pre-tested and validated questionnaire. The list of 8 variables of study and their levels is shown below:

fP: Use of modern family planning methods (1, 2)
res: Residential area of respondent (1, 2)
sti: Experience of having STI's (1, 2)
famsize: Family size of respondent (1, 2)
marital: Marital status of respondent (1, 2)
vct: Voluntary counselling and testing for HIV (1, 2)
educ: Level of education of respondent (1, 2, 3, 4)
age: Age of respondent in years (# #)

Summary statistics for the variable age in years

. use f:\fp1.dta, clear

. su age, detail

```
                            age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           18             18
 5%           18             18
10%           20             18       Obs                 474
25%           24             18       Sum of Wgt.         474

50%           35                      Mean           38.16667
```

```
                        Largest       Std. Dev.      16.37521
75%             51             85
90%             60             85       Variance       268.1476
95%             68             87       Skewness       .6828318
99%             82             89       Kurtosis       2.629198
```

Frequency tables for discrete variables

```
. use f:\fp1.dta, clear

. tab1   fp res sti famsize marital vct educ

-> tabulation of fp

         fp |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |        156       32.91       32.91
        Yes |        318       67.09      100.00
------------+-----------------------------------
      Total |        474      100.00

-> tabulation of res

        res |      Freq.     Percent        Cum.
------------+-----------------------------------
      Rural |        328       69.20       69.20
      Urban |        146       30.80      100.00
------------+-----------------------------------
      Total |        474      100.00

-> tabulation of sti

        sti |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |         60       12.66       12.66
        Yes |        414       87.34      100.00
------------+-----------------------------------
      Total |        474      100.00

-> tabulation of famsize

    famsize |      Freq.     Percent        Cum.
------------+-----------------------------------
  5 or less |        370       78.06       78.06
More than 5 |        104       21.94      100.00
------------+-----------------------------------
      Total |        474      100.00

-> tabulation of marital

     marital |      Freq.     Percent        Cum.
```

Married	258	54.43	54.43
Not married	216	45.57	100.00
Total	474	100.00	

-> tabulation of vct

vct	Freq.	Percent	Cum.
Not willing	84	17.72	17.72
Tested	363	76.58	94.30
Willing	27	5.70	100.00
Total	474	100.00	

-> tabulation of educ

educ	Freq.	Percent	Cum.
Postgraduate	59	12.45	12.45
Primary or less	109	23.00	35.44
Secondary	190	40.08	75.53
Undergraduate	116	24.47	100.00
Total	474	100.00	

Pearson's chi-square test of association between fp and vct

. tab2 vct fp, cell exp chi2

-> tabulation of vct by fp

Key
frequency
expected frequency
cell percentage

vct	fp No	fp Yes	Total
Not willing	23	61	84
	27.6	56.4	84.0
	4.85	12.87	17.72

```
-------------+----------------------+----------
      Tested |       106        257 |       363
             |     119.5      243.5 |     363.0
             |     22.36      54.22 |     76.58
-------------+----------------------+----------
     Willing |        27           0 |        27
             |       8.9        18.1 |      27.0
             |      5.70        0.00 |      5.70
-------------+----------------------+----------
       Total |       156        318 |       474
             |     156.0      318.0 |     474.0
             |     32.91      67.09 |    100.00

          Pearson chi2(2) =  58.4653   Pr = 0.000
```

How to obtain crude odds and risk ratios

For the relationship between the variables **res** and **fp**, estimate crude odds and risk ratios and 95% confidence intervals.

. tab2 res fp

-> tabulation of res by fp

```
           |          fp
       res |        No        Yes |     Total
-----------+----------------------+----------
     Rural |       119        209 |       328
     Urban |        37        109 |       146
-----------+----------------------+----------
     Total |       156        318 |       474
```

. csi 119 209 37 109, or

```
                |   Exposed   Unexposed  |      Total
----------------+----------------------+------------
          Cases |       119         209  |        328
       Noncases |        37         109  |        146
----------------+----------------------+------------
          Total |       156         318  |        474
                |                        |
           Risk |  .7628205    .6572327  |   .6919831
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
Risk difference |         .1055878       |    .0208729    .1903027
```

```
      Risk ratio |         1.160655       |    1.031327      1.3062
 Attr. frac. ex. |         .1384176       |    .0303758    .2344207
 Attr. frac. pop |         .0502186       |
      Odds ratio |         1.677357       |     1.08686    2.587951
(Cornfield)
                 +-------------------------------------------------
                                chi2(1) =     5.47  Pr>chi2 = 0.0193
```

Perform the Pearson chi-square test of association between **educ** and **fp** at the 0.05 level of significance.

. tab2 educ fp, cell exp chi2

```
-> tabulation of educ by fp

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|  cell percentage   |
+--------------------+

                |          fp
           educ |        No        Yes |     Total
----------------+----------------------+----------
   Postgraduate |        14         45 |        59
                |      19.4       39.6 |      59.0
                |      2.95       9.49 |     12.45
----------------+----------------------+----------
Primary or less |        48         61 |       109
                |      35.9       73.1 |     109.0
                |     10.13      12.87 |     23.00
----------------+----------------------+----------
      Secondary |        70        120 |       190
                |      62.5      127.5 |     190.0
                |     14.77      25.32 |     40.08
----------------+----------------------+----------
  Undergraduate |        24         92 |       116
                |      38.2       77.8 |     116.0
                |      5.06      19.41 |     24.47
----------------+----------------------+----------
          Total |       156        318 |       474
                |     156.0      318.0 |     474.0
                |     32.91      67.09 |    100.00

          Pearson chi2(3) =  17.5403   Pr = 0.001
```

How to perform several Pearson's chi-square tests of association

Perform the Pearson chi-square test of association between **fp** and each of the variables **res**, **sti**, **famsize**,

marital, vct and educ at the 0.05 level of significance.

. for var vct res educ sti marital skills perception famsize: tab2 fp X, cell exp chi2

```
->  tab2 fp res, cell exp chi2

-> tabulation of fp by res

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|  cell percentage   |
+--------------------+

           |          res
        fp |     Rural      Urban |     Total
-----------+----------------------+----------
        No |       119         37 |       156
           |     107.9       48.1 |     156.0
           |     25.11       7.81 |     32.91
-----------+----------------------+----------
       Yes |       209        109 |       318
           |     220.1       97.9 |     318.0
           |     44.09      23.00 |     67.09
-----------+----------------------+----------
     Total |       328        146 |       474
           |     328.0      146.0 |     474.0
           |     69.20      30.80 |    100.00

          Pearson chi2(1) =   5.4743   Pr = 0.019

->  tab2 fp sti, cell exp chi2

-> tabulation of fp by sti

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|  cell percentage   |
+--------------------+

           |          sti
        fp |        No        Yes |     Total
-----------+----------------------+----------
        No |        22        134 |       156
           |      19.7      136.3 |     156.0
           |      4.64      28.27 |     32.91
-----------+----------------------+----------
       Yes |        38        280 |       318
           |      40.3      277.7 |     318.0
           |      8.02      59.07 |     67.09
-----------+----------------------+----------
     Total |        60        414 |       474
           |      60.0      414.0 |     474.0
           |     12.66      87.34 |    100.00

          Pearson chi2(1) =   0.4388   Pr = 0.508

->  tab2 fp famsize, cell exp chi2

-> tabulation of fp by famsize

+--------------------+
| Key                |
```

```
|--------------------|
|     frequency      |
| expected frequency |
|  cell percentage   |
+--------------------+

           |        famsize
        fp | 5 or less  More than |     Total
-----------+----------------------+----------
        No |       101         55 |       156
           |     121.8       34.2 |     156.0
           |     21.31      11.60 |     32.91
-----------+----------------------+----------
       Yes |       269         49 |       318
           |     248.2       69.8 |     318.0
           |     56.75      10.34 |     67.09
-----------+----------------------+----------
     Total |       370        104 |       474
           |     370.0      104.0 |     474.0
           |     78.06      21.94 |    100.00

          Pearson chi2(1) =  24.0719   Pr = 0.000

->  tab2 fp marital, cell exp chi2

-> tabulation of fp by marital

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|  cell percentage   |
+--------------------+

           |        marital
        fp |   Married  Not marri |     Total
-----------+----------------------+----------
        No |        93         63 |       156
           |      84.9       71.1 |     156.0
           |     19.62      13.29 |     32.91
-----------+----------------------+----------
       Yes |       165        153 |       318
           |     173.1      144.9 |     318.0
           |     34.81      32.28 |     67.09
-----------+----------------------+----------
     Total |       258        216 |       474
           |     258.0      216.0 |     474.0
           |     54.43      45.57 |    100.00

          Pearson chi2(1) =   2.5203   Pr = 0.112

->  tab2 fp vct, cell exp chi2

-> tabulation of fp by vct

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|  cell percentage   |
+--------------------+

           |               vct
        fp | Not willi     Tested    Willing |     Total
-----------+---------------------------------+----------
        No |        23        106         27 |       156
           |      27.6      119.5        8.9 |     156.0
           |      4.85      22.36       5.70 |     32.91
-----------+---------------------------------+----------
       Yes |        61        257          0 |       318
           |      56.4      243.5       18.1 |     318.0
           |     12.87      54.22       0.00 |     67.09
-----------+---------------------------------+----------
     Total |        84        363         27 |       474
           |      84.0      363.0       27.0 |     474.0
```

```
           |     17.72      76.58       5.70 |    100.00

          Pearson chi2(2) =  58.4653   Pr = 0.000

->  tab2 fp educ, cell exp chi2

-> tabulation of fp by educ

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|  cell percentage   |
+--------------------+

           |                    educ
        fp | Postgradu  Primary o  Secondary  Undergrad |     Total
-----------+--------------------------------------------+----------
        No |        14          48          70         24 |       156
           |      19.4        35.9        62.5       38.2 |     156.0
           |      2.95       10.13       14.77       5.06 |     32.91
-----------+--------------------------------------------+----------
       Yes |        45          61         120         92 |       318
           |      39.6        73.1       127.5       77.8 |     318.0
           |      9.49       12.87       25.32      19.41 |     67.09
-----------+--------------------------------------------+----------
     Total |        59         109         190        116 |       474
           |      59.0       109.0       190.0       116.0 |     474.0
           |     12.45       23.00       40.08       24.47 |    100.00

          Pearson chi2(3) =  17.5403   Pr = 0.001
```

How to convert a continuous variable into a categorical variable

Open up the data set **fp1.dta** in STATA. Create the following categorical variables out of the continuous variables you have in the data set fp1.dta. Save your newly created STATA data set as **fp2.dta**.

fp: (1 if no fp used, 0 otherwise)
res: (1 if rural, 0 if urban)
sti: (1 if yes, 0 if no)
famsize: (1 if more than 5, 0 if 5 or less)
marital: (1 if not married, 0 if married)
vct: (1 if tested, 2 if willing, 3 if not willing)
educ: (1 if postgraduate, 2 if undergraduate, 3 if secondary, 4 if primary or less)
age: Age in years
agecat: (1, 2, 3, 4)

a) Break down the variable **age** into 4 categories in STATA, call it **agecat**, and find out if there is a significant association between the variables **fp** and **agecat** at the 0.05 level.

The categories of variable **agecat** are as follows:

Age category 1: Ages less than 24
Age category 2: Ages between 24 (inclusive) and 35
Age category 3: Ages between 35 (inclusive) and 51
Age category 4: Ages greater than or equal to 51

$$\text{agecat} = \begin{cases} 1 \ if \ \ age < 24 \\ 2 \ if \ \ age \in [24,\ 35) \\ 3 \ if \ \ age \in [35,\ 51) \\ 4 \ if \ \ age \geq 51 \end{cases}$$

b) The outcome variable of study is **fp**. Identify factors that affect **fp** significantly based on Pearson's chi-square tests of association at the 0.05 level of significance.

c) Repeat the exercise in (b) by using binary logistic regression analysis. Assess the reliability of your fitted logistic regression model.

d) Repeat the exercise in (c) by treating the variables vct, educ and agecat as categorical variables. Assess the reliability of your fitted logistic regression model.

Solutions

Step 1: Create the variable agecat as follows:

```
. use f:\fp1.dta, clear

. gen agecat=.
(474 missing values generated)

. replace agecat=1 if age < 24
(114 real changes made)

. replace agecat=2 if age >= 24  &  age < 35
(122 real changes made)
```

```
. replace agecat=3 if age >= 35  &  age < 51
(119 real changes made)

. replace agecat=4 if age >= 51
(119 real changes made)

. tab agecat
```

agecat	Freq.	Percent	Cum.
1	114	24.05	24.05
2	122	25.74	49.79
3	119	25.11	74.89
4	119	25.11	100.00
Total	474	100.00	

Binary logistic regression analysis (Treating vct, educ and agecat as dichotomous variables)

. use f:\fp2.dta, clear

. logistic fp res sti famsize marital vct educ agecat

```
Logistic regression                               Number of obs   =        474
                                                  LR chi2(7)      =      55.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -272.75864                       Pseudo R2       =     0.0917
```

fp	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
res	2.956097	1.172459	2.73	0.006	1.358663	6.431698
sti	1.539002	.8472669	0.78	0.434	.5231513	4.527425
famsize	2.960315	.7314342	4.39	0.000	1.823993	4.80455
marital	1.380184	.325375	1.37	0.172	.8694972	2.190815
vct	2.037644	.5184965	2.80	0.005	1.237464	3.355244
educ	1.438571	.1877199	2.79	0.005	1.113929	1.857826
agecat	.8250634	.0784518	-2.02	0.043	.6847785	.9940872

To obtain the classification table

. lstat

```
Logistic model for fp

             -------- True --------
Classified |         D            ~D  |      Total
```

```
-----------+--------------------------+-----------
     +     |        45            26  |         71
     -     |       111           292  |        403
-----------+--------------------------+-----------
   Total   |       156           318  |        474

Classified + if predicted Pr(D) >= .5
True D defined as fp != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   28.85%
Specificity                     Pr( -|~D)   91.82%
Positive predictive value       Pr( D| +)   63.38%
Negative predictive value       Pr(~D| -)   72.46%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.18%
False - rate for true D         Pr( -| D)   71.15%
False + rate for classified +   Pr(~D| +)   36.62%
False - rate for classified -   Pr( D| -)   27.54%
--------------------------------------------------
Correctly classified                        71.10%
--------------------------------------------------
```

The percentage of overall correct classification is 71.10% < 75%. The fitted model is poorly sensitive (28.85% < 50%) and highly specific (91.82% > 50%). The fitted model is **not** reliable.

The Hosmer-Lemeshow goodness of fit test

. lfit

```
Logistic model for fp, goodness-of-fit test

      number of observations =       474
 number of covariate patterns =       117
           Pearson chi2(109) =       176.68
                 Prob > chi2 =         0.0000
```

Since the P-value is 0.0000 < 0.05, the fitted model is not reliable.

Binary logistic regression analysis (Treating vct, educ and agecat as categorical variables)

. xi: logistic fp res sti famsize marital i.vct i.educ i.agecat

```
i.vct             _Ivct_1-3           (naturally coded; _Ivct_1 omitted)
i.educ            _Ieduc_1-4          (naturally coded; _Ieduc_1 omitted)
i.agecat          _Iagecat_1-4        (naturally coded; _Iagecat_1 omitted)

note: _Ivct_2 != 0 predicts success perfectly
      _Ivct_2 dropped and 27 obs not used

Logistic regression                               Number of obs   =        447
```

```
                                                LR chi2(11)     =      37.67
                                                Prob > chi2     =     0.0001
Log likelihood = -249.76158                     Pseudo R2       =     0.0701

------------------------------------------------------------------------------
          fp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         res |   2.182425   .9240777     1.84   0.065     .9517554    5.004414
         sti |   .1737471   .1274216    -2.39   0.017     .0412732      .73142
     famsize |   2.669966   .7013619     3.74   0.000     1.595538    4.467908
     marital |   1.101077   .2925194     0.36   0.717      .654157    1.853333
     _Ivct_3 |   .3573932   .2579187    -1.43   0.154     .0868686    1.470381
    _Ieduc_2 |    .377254   .1809954    -2.03   0.042     .1473168    .9660851
    _Ieduc_3 |   .6549182   .2992938    -0.93   0.354     .2674194    1.603914
    _Ieduc_4 |   .8751554   .3929063    -0.30   0.766     .3630221     2.10978
  _Iagecat_2 |   .8213294   .2486699    -0.65   0.516      .453736    1.486728
  _Iagecat_3 |   .6969106   .2188487    -1.15   0.250      .376597    1.289666
  _Iagecat_4 |   .5650128   .1786612    -1.81   0.071       .30402    1.050061
------------------------------------------------------------------------------
```

How to detect a confounding variable

Consider the binary logistic regression of variable **fp** on the 7 dichotomous variables. Find out if the variable **agecat** is a confounding variable.

First method: Using the tabodds command in STATA

. use f:\fp2.dta, clear

. tabodds fp res, or

```
------------------------------------------------------------------------------
        res |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
------------+-----------------------------------------------------------------
          0 |    1.000000          .            .            .           .
          1 |    1.677357       5.46       0.0194     1.081817    2.600741
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     5.46
                                  Pr>chi2  =   0.0194

Score test for trend of odds:     chi2(1)  =     5.46
                                  Pr>chi2  =   0.0194
```

. tabodds fp res, or adjust(agecat)

Mantel-Haenszel odds ratios adjusted for agecat

```
------------------------------------------------------------------------------
        res |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
------------+-----------------------------------------------------------------
          0 |    1.000000          .            .            .           .
          1 |    1.733317       6.17       0.0130     1.117065    2.689538
------------------------------------------------------------------------------
Score test for trend of odds: chi2(1)  =    6.17
                              Pr>chi2  =  0.0130
```

The 2 P-values have not changed. Hence, agecat is not a confounding variable.

. tabodds fp famsize, or

```
------------------------------------------------------------------------------
     famsize |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |    1.000000          .            .            .           .
           1 |    2.989493      24.02       0.0000     1.886944    4.736264
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =    24.02
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    24.02
                                  Pr>chi2  =   0.0000
```

. tabodds fp famsize, or adjust(agecat)

```
Mantel-Haenszel odds ratios adjusted for agecat

------------------------------------------------------------------------------
     famsize |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |    1.000000          .            .            .           .
           1 |    3.181150      26.10       0.0000     1.989887    5.085573
------------------------------------------------------------------------------
Score test for trend of odds: chi2(1)  =   26.10
                              Pr>chi2  =  0.0000
```

The 2 P-values have not changed. Hence, agecat is not a confounding variable.

Second method: Using the adjust procedure in STATA

. use f:\fp2.dta, clear

. logistic fp res sti famsize marital vct educ agecat

```
Logistic regression                               Number of obs   =        474
                                                  LR chi2(7)      =      55.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -272.75864                       Pseudo R2       =     0.0917

------------------------------------------------------------------------------
          fp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         res |   2.956097   1.172459     2.73   0.006     1.358663    6.431698
         sti |   1.539002   .8472669     0.78   0.434     .5231513    4.527425
     famsize |   2.960315   .7314342     4.39   0.000     1.823993     4.80455
     marital |   1.380184    .325375     1.37   0.172     .8694972    2.190815
         vct |   2.037644   .5184965     2.80   0.005     1.237464    3.355244
        educ |   1.438571   .1877199     2.79   0.005     1.113929    1.857826
      agecat |   .8250634   .0784518    -2.02   0.043     .6847785    .9940872
------------------------------------------------------------------------------
```

```
. adjust   agecat,  by (res  famsize)  exp  ci

------------------------------------------------------------
     Dependent variable: fp     Command: logistic
   Variables left as is: sti, marital, vct, educ
  Covariate set to mean: agecat = 2.5126582
------------------------------------------------------------

-----------------------------------------------
          |                famsize
      res |                  0                1
----------+------------------------------------
        0 |             .25212          .623558
          | [.165226,.384712]  [.329074,1.18157]
          |
        1 |           .400746           1.30469
          |  [.30288,.530236]    [.8667,1.96402]
-----------------------------------------------
     Key:  exp(xb)
           [95% Confidence Interval]
```

STATA produces adjusted odds ratios for the variables res (1, 0) and famsize (1, 0) by calculating estimates for all possible combinations between the possible values of res and famsize. We have **2 X 2 = 4 possible combinations** between the categories of the 2 variables for which adjusted odds ratios are required.

The adjusted odds ratio for res corresponds to res=1 and famsize=0. It is equal to exp[0.400746], with a 95% confidence interval of [exp(0.31), exp(0.53)]. That is, the adjusted odds ratio for **res** is equal to **1.49**, with a 95% confidence interval of **[1.35, 1.70]**.

The adjusted odds ratio for famsize corresponds to famsize=1 and res=0. It is equal to exp[0.623558], with a 95% confidence interval of [exp(0.33), exp(1.18)]. That is, the adjusted odds ratio for **famsize** is equal to **1.87**, with a 95% confidence interval of **[1.39, 3.26]**.

Hence, we have the following summary table of unadjusted and adjusted odds ratios for the 2 risk factors (water and residence):

Table 1: Unadjusted and *adjusted odds ratios for res and famsize

Predictor variable	Unadjusted odds ratios and 95% C. I.	*Adjusted odds ratios and 95% C. I.
res	2.96 (1.36, 6.43)	1.49 (1.35, 1.70)
famsize	2.96 (1.83, 4.81)	1.87 (1.39, 3.26)

***Adjustment was done for age category**

The adjusted and unadjusted odds ratios do not differ significantly from each other. This is because both sets of confidence intervals exclude 1. Hence, the variable **agecat** is not a confounding variable.