Notes for laboratory session 6

Analysis using a 2×2 table or logistic regression

Factor with two levels (“`more`”) .

Consider the 2×2 table tabulating the use of contraceptives among women that desire more children, versus women that want no more children:

```
. tabulate cuse more [freq=N], chi

Contracept |     Desires more
   ive use |       children?
  (Yes/No) |        No        Yes |     Total
-----------+----------------------+----------
        No |       347        753 |      1100
       Yes |       288        219 |       507
-----------+----------------------+----------
     Total |       635        972 |      1607

          Pearson chi2(1) =  92.6442   Pr = 0.000
```

a) Calculate the p-value for the chi-square statistic using the appropriate STATA function.

Using STATA `logit` command this analysis looks as follows (note that we use “No use” as the reference cell). The likelihood of this model is saved with the `lrtest` command:

```
. char more[omit] 0
. xi: logit cuse i.more [freq=N], nolog
i.more                Imore_0-1    (naturally coded; Imore_0 omitted)

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(1)      =      91.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -956.00957                       Pseudo R2       =     0.0458

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Imore_1 |  -1.048629    .110672     -9.475   0.000      -1.265542    -.831716
   _cons |  -.1863643   .0797124     -2.338   0.019      -.3425977   -.0301309
------------------------------------------------------------------------------

. est store M1
```

b) Compare the chi-square statistic in the `logit` command output with the one given in 2×2 table analysis.
c) Calculate the Odds for the use of contraceptives in the two “`more`” categories.
d) Calculate the Odds Ratio. Now use the 2×2 table data to produce the Odds Ratio. Compare the two OR’s.
e) How can we test the significance of the “`more`” predictor? How is the relevant statistic produced? What are the distributional properties of this statistic?

Produce estimates of the odds ratios

i. By including the option `or` after the logit statement, or

ii. By using the `logistic` command.

```
. logit , or

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(1)      =      91.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -956.00957                       Pseudo R2       =     0.0458

------------------------------------------------------------------------------
    cuse | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    more |   .3504178   .0387814     -9.475   0.000       .2820863    .4353017


. xi: logistic cuse i. more [freq=N]

i.more                Imore_0-1    (naturally coded; Imore_0 omitted)

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(1)      =      91.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -956.00957                       Pseudo R2       =     0.0458

------------------------------------------------------------------------------
    cuse | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    more |   .3504178   .0387814     -9.475   0.000       .2820863    .4353017
```

f) How is the 95% Confidence Interval for the OR produced in the `logistic` command output?

<u>The "null" model</u>

Consider the following model:

```
. xi: logit cuse [freq=N], nolog

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -1001.8468                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   _cons |  -.7745545   .0536794    -14.429   0.000      -.8797641   -.6693448
------------------------------------------------------------------------------
. est store M0
```

g) What is the interpretation of the β_0 coefficient? Check your result using the 2×2 table data.

h) Calculate the –2logλ statistic using the maximized likelihoods in the null model and the model with the "`more`" predictor. Compare your result with the z-statistic for the variable "`more`".

The (Wald) chi-square statistic can be obtained by the test command in STATA as follows:

```
. test Imore_1

 ( 1)  Imore_1 = 0.0

           chi2(  1) =   89.78
         Prob > chi2 =    0.0000
```

Analysis using a 2×c table or logistic regression

<u>Factor with more than two levels ("`age`") .</u>

Consider the 2×4 table tabulating the use of contraceptives among four different age groups:

```
Contracept |
   ive use |                    Age
  (Yes/No) |       <25      25-29      30-39      40-49 |     Total
-----------+--------------------------------------------+----------
        No |       325        299        375        101 |      1100
       Yes |        72        105        237         93 |       507
-----------+--------------------------------------------+----------
     Total |       397        404        612        194 |      1607
```

Using STATA logit command the same analysis looks like follows:

```
. char age[omit] 1
. xi: logit cuse i.age [freq=N] nolog
i.age                 Iage_1-4     (naturally coded; Iage_1 omitted)

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(3)      =      79.19
                                                  Prob > chi2     =     0.0000
Log likelihood = -962.25091                       Pseudo R2       =     0.0395

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Iage_2 |   .4606758   .1727254      2.667   0.008       .7992114    .1221403
  Iage_3 |   1.048293   .1544404      6.788   0.000       1.350991    .7455955
  Iage_4 |   1.424638   .1939573      7.345   0.000       1.804787    1.044489
   _cons |  -1.507159   .1302527    -11.571   0.000      -1.251868    -1.76245

. est store M2
```

a) What is the value of the likelihood ratio statistic? Compare it to the appropriate distribution in order to obtain the relevant p-value.
b) Calculate the odds ratios of each age group compared to the reference group. Derive now the same Odss Ratios using the 2×4 table and compare the two approaches.
c) How can we check the significance of each group individually? Do you notice any kind of pattern in the age group coefficients.

We can test the significance of the age factor globally, using a Wald chi-square test:

```
. test Iage_2 Iage_3 Iage_4

 ( 1)  Iage_2 = 0.0
 ( 2)  Iage_3 = 0.0
 ( 3)  Iage_4 = 0.0

           chi2(  3) =   74.36
         Prob > chi2 =    0.0000
```

Two factors

Suppose that we introduce in the model both factors age and more. The tables are broken down by age as follows:

```
. sort age

. by age: tab cuse more [freq=N]
```

Or in a more compact way

```
. bysort age: tab cuse more [freq=N]
```

```
-> age=      <25
Contracept |     Desires more
   ive use |       children?
  (Yes/No) |       Yes         No |     Total
-----------+----------------------+----------
        No |       265         60 |       325
       Yes |        58         14 |        72
-----------+----------------------+----------
     Total |       323         74 |       397
```

```
-> age=    25-29
Contracept |     Desires more
   ive use |       children?
  (Yes/No) |       Yes         No |     Total
-----------+----------------------+----------
        No |       215         84 |       299
       Yes |        68         37 |       105
-----------+----------------------+----------
     Total |       283        121 |       404
```

```
-> age=    30-39
Contracept |     Desires more
   ive use |       children?
  (Yes/No) |       Yes         No |     Total
-----------+----------------------+----------
        No |       230        145 |       375
       Yes |        79        158 |       237
-----------+----------------------+----------
     Total |       309        303 |       612
```

```
-> age=    40-49
Contracept |     Desires more
   ive use |       children?
  (Yes/No) |       Yes         No |     Total
-----------+----------------------+----------
        No |        43         58 |       101
       Yes |        14         79 |        93
-----------+----------------------+----------
     Total |        57        137 |       194
```

Use the Mantel-Haenszel (M-H) analysis to adjust for age the relationship of contraceptive use and desire for more children.

```
. cc cuse more [freq=N], by(age)

              Age |       OR      [95% Conf. Interval]    M-H Weight
 -----------------+-------------------------------------------------
              <25 |   .9380054      .4944402   1.776932      9.345088 (Cornfield)
            25-29 |    .718039      .4481752   1.150032      19.69059 (Cornfield)
            30-39 |   .3152174       .224304   .4429905      59.37908 (Cornfield)
            40-49 |   .2390344      .1206217   .4744326      17.51031 (Cornfield)
 -----------------+-------------------------------------------------
            Crude |   .3504178      .2821249   .4352413                (Cornfield)
     M-H combined |   .4324495      .3432378   .5448483
 -----------------+-------------------------------------------------
 Test of homogeneity (M-H)      chi2(3) =    16.03  Pr>chi2 = 0.0011

                    Test that combined OR = 1:
                                 Mantel-Haenszel chi2(1) =     50.36
                                                 Pr>chi2 =    0.0000
```

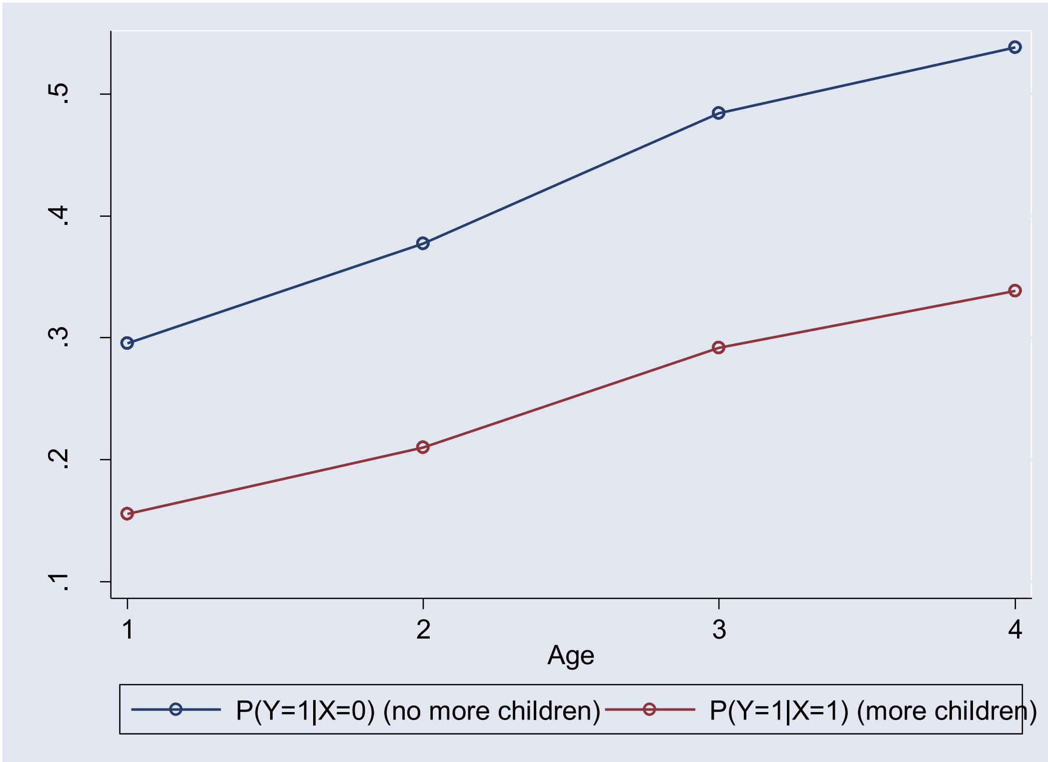
a) Is the relationship between contraceptive use and desire for more children significant?
b) The test for homogeneity is significant. What is the interpretation of this result?

A more flexible way to proceed is via logistic regression models:

```
. xi: logit cuse i.age i.more [freq=N] ,nolog

   i.age                Iage_1-4     (naturally coded; Iage_1 omitted)
   i.more               Imore_0-1    (naturally coded; Imore_0 omitted)

   Logit estimates                                   Number of obs   =       1607
                                                     LR chi2(4)      =     128.88
                                                     Prob > chi2     =     0.0000
   Log likelihood = -937.40449                       Pseudo R2       =     0.0643

   ------------------------------------------------------------------------------
       cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
   ---------+--------------------------------------------------------------------
     Iage_2 |   .3678306   .1753673      2.097   0.036        .024117    .7115443
     Iage_3 |   .8077888   .1597533      5.056   0.000        .494678      1.1209
     Iage_4 |   1.022618   .2039337      5.014   0.000       .6229158    1.422321
    Imore_1 |   -.824092   .1171128     -7.037   0.000      -1.053629   -.5945552
      _cons |  -.8698414   .1571298     -5.536   0.000       -1.17781   -.5618727

   . est store M3
```

The above model is shown graphically as follows:

```
. quietly xi: logit cuse i.age more [freq=N]
. predict phat
(option p assumed; Pr(cuse))
. generate phat0=phat if more==0
. generate phat1=phat if more==1
. label var phat0 "P(Y=1|X=0) (no more children)"
. label var phat1 "P(Y=1|X=1) (more children)"
. sort age
. sc phat0 phat1 age, xlab() ylab()  l1(Probability) c(l l)
```



c) Try to produce a similar graph for the log(Odds) instead of probabilities. (Check the STATA help file for the `logistic` command in order to locate the appropriate option for the `predict` command)

d) Calculate the adjusted for age estimate of the odds ratio of using contraception , associated with the desire for more children versus desire for no more children.

e) Calculate the adjusted for desire for more children estimate of the odds ratio of using contraception versus not using for women aged 40-49 vs. women aged <25.

f) What is the underlying assumption of the previous model about the difference between the two "`more`" groups across the four age group categories.

The two-factor model with interaction

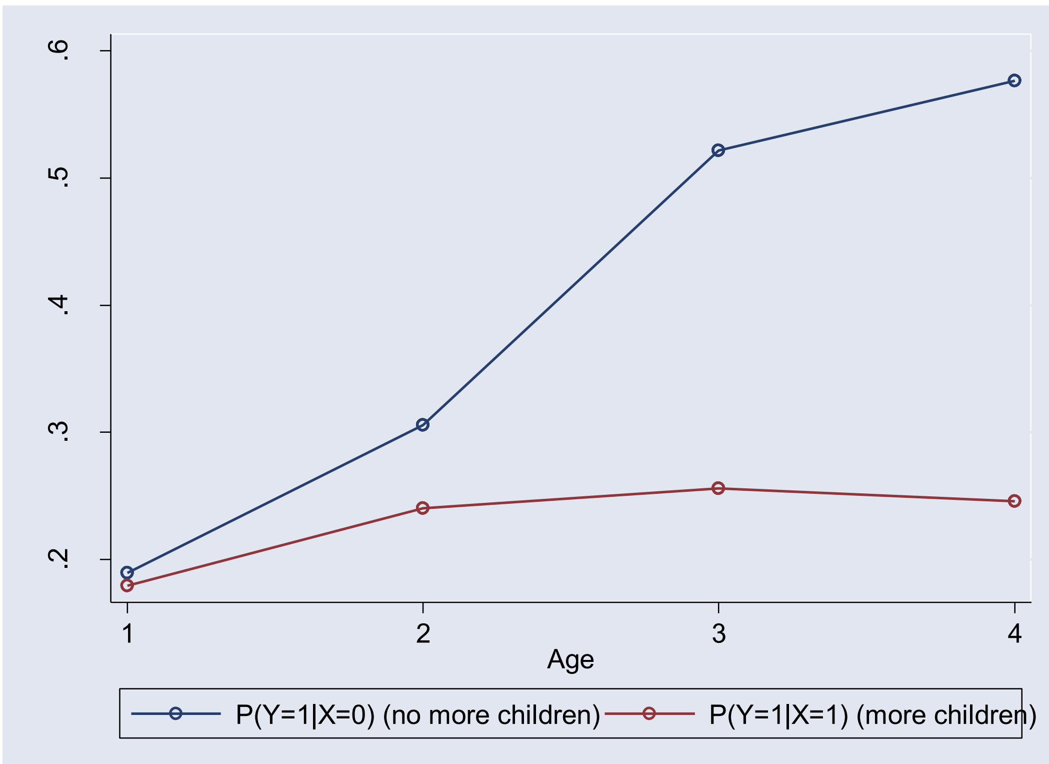
Consider the previous logistic regression model with the addition of the `more-age` interaction.

```
. xi: logit cuse i.age i.more i.age*i.more  [freq=N],nolog
i.age                Iage_1-4     (naturally coded; Iage_1 omitted)
i.more               Imore_0-1    (naturally coded; Imore_0 omitted)
i.age*i.more         IaXm_#-#     (coded as above)
Note: Iage_2 dropped due to collinearity.
Note: Iage_3 dropped due to collinearity.
Note: Iage_4 dropped due to collinearity.
Note: Imore_1 dropped due to collinearity.

Logit estimates                                   Number of obs   =       1607
                                                 LR chi2(7)      =     145.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -929.01009                       Pseudo R2       =     0.0727

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Iage_2 |   .6353883   .3564083      1.783   0.075      -.0631592    1.333936
  Iage_3 |   1.541149   .3183093      4.842   0.000       .9172739    2.165023
  Iage_4 |   1.764292   .3435036      5.136   0.000       1.091037    2.437547
 Imore_1 |  -.0639996    .330318     -0.194   0.846       -.711411    .5834119
IaXm_2_1 |  -.2672319    .409144     -0.653   0.514      -1.069139    .5346757
IaXm_3_1 |  -1.090493    .373285     -2.921   0.003      -1.822118   -.3588679
IaXm_4_1 |  -1.367148   .4834191     -2.828   0.005      -2.314632   -.4196641
   _cons |  -1.455287   .2968082     -4.903   0.000      -2.037021   -.8735538
------------------------------------------------------------------------------
.est store M4
```

a) Calculate the adjusted estimate of the odds ratio of using contraception versus not using for women aged 40-49 vs. women aged <25 i. For women desiring more children and ii. For women not desiring more children. What is the interpretation of the interaction term (`IaXm_4_1`) coefficient.
b) What is the main difference between the models with and without the interaction term?

Graphically, the model with interaction can be shown as follows:

```
. predict phatx
(option p assumed; Pr(cuse))
. gen phatx0=phatx if more==0
(16 missing values generated)
. gen phatx1=phatx if more==1
(16 missing values generated)
. label var phatx1 "P(Y=1|X=1) (more children)"
. label var phatx0 "P(Y=1|X=0) (no more children)"
. sort age
. sc phatx0 phatx1 age , xlab() ylab() c(l l) l1(Probability)
```



c) Produce a similar graph showing Odds instead of probabilities.

Model selection

The best model can be determined by considering the likelihood-ratio statistics produced in the STATA output above:

1. Model with more versus the null model

```
. lrtest M0 M1
likelihood-ratio test                                 LR chi2(1)  =     91.67

(Assumption: M0 nested in M1)                         Prob > chi2 =    0.0000
```

2. Model with age versus the null model

```
. lrtest M0 M2


likelihood-ratio test                                 LR chi2(3)  =     79.19

(Assumption: M0 nested in M2)                         Prob > chi2 =    0.0000
```

3. Model with more versus the two-factor model with no interaction

```
. lrtest M1 M3


likelihood-ratio test                                 LR chi2(3)  =     37.21

(Assumption: M1 nested in M3)                         Prob > chi2 =    0.0000
```

4. Model with age versus the two-factor model with no interaction

```
. lrtest M2 M3


likelihood-ratio test                                 LR chi2(1)  =     49.69

(Assumption: M2 nested in M3)                         Prob > chi2 =    0.0000
```

5. The effect of interaction is given from the following test:

```
. lrtest M3 M4


likelihood-ratio test                                 LR chi2(3)  =     16.79

(Assumption: M3 nested in M4)                         Prob > chi2 =    0.0008
```

Fill the following table. P_{n+1} is the "smaller" model which is nested in the previous model P_n and l is the maximized log likelihood.

Model	**Log Likelihood (l)**	**$-2*[l(P_{n+1})-l(P_n)]$**	**Df**	**p-value**
Two factors (with interaction)		—	-	—
Two factors (no interaction)				
Age				
Desires more children?				
Null model				

a) What do conclude about the significance of the interaction term?

Analysis of covariance-type models

Given the strong linear relationship between the logit of contraceptive use and age, we may consider a model where age is not grouped in categories but is entered as a continuous covariate.

```
. gen contage = age

. recode contage 1=20 2=27.5 3=35 4=45
(32 changes made)
```

Single-factor model

The single-factor model is given as follows:

```
. logit cuse contage [freq=N], nolog

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(1)      =      76.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -963.45258                       Pseudo R2       =     0.0383

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 contage |    .060671   .0071034     -8.541   0.000      -.0745934   -.0467486
   _cons |  -2.672667   .2332492     11.458   0.000       2.215507    3.129827

.est store M5
```

a) What is the interpretation of the "`contage`" coefficient?
b) What is the main advantage of this approach instead of the previous age parametrization? What is the differnece in our assumptions when we use age as a continuous variable?

Two-factor model with no interaction

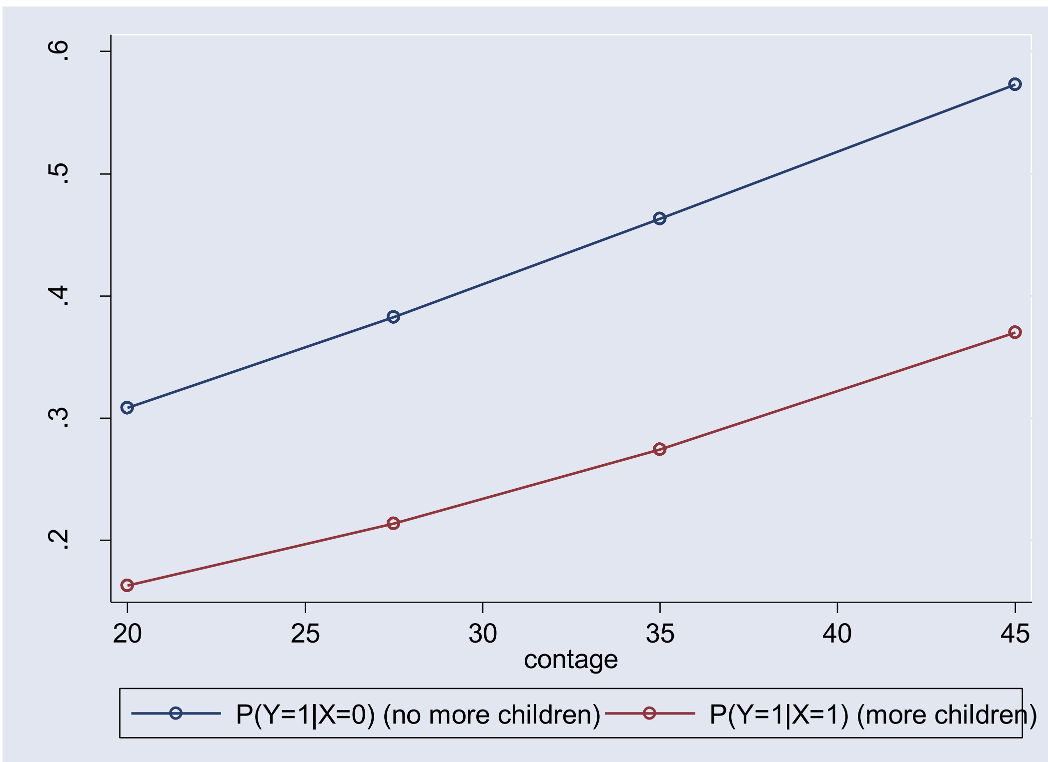
The model including both age and desire for more children is given as follows:

```
. xi: logit cuse i.more contage [freq=N], nolog
i.more              Imore_0-1    (naturally coded; Imore_0 omitted)

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(2)      =     126.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -938.50406                       Pseudo R2       =     0.0632

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Imore_1 |  -.8258978     .11711     -7.052   0.000      -1.055429   -.5963665
 contage |  -.0441062    .007529     -5.858   0.000      -.0588627   -.0293497
   _cons |   2.516654   .2365292     10.640   0.000       2.053065    2.980243
------------------------------------------------------------------------------
. est store M6
. lrtest M5 M6

likelihood-ratio test                                  LR chi2(1)  =     49.90
(Assumption: M5 nested in M6)                          Prob > chi2 =    0.0000
```

c) Is the effect of the "`more`" variable significant? Notice the relation between the chi-square statistic in the `lrtest` output and the z-statistic for the "`more`" variable in the `logit` command output.

Graphically, the model with interaction can be shown as follows:

```
. predict yhat
(option p assumed; Pr(cuse))
. generate yhat1=yhat if more==1
(16 missing values generated)
. generate yhat0=yhat if more==0
(16 missing values generated)
. label var yhat1 "P(Y=1|X=1) (more children)"
. label var yhat0 "P(Y=1|X=0) (no more children)"
. sort more age
. sc yhat0 yhat1 contage, c(l l) xlab() ylab() l1("Probability")
```



d) Why are the lines not exactly straight?

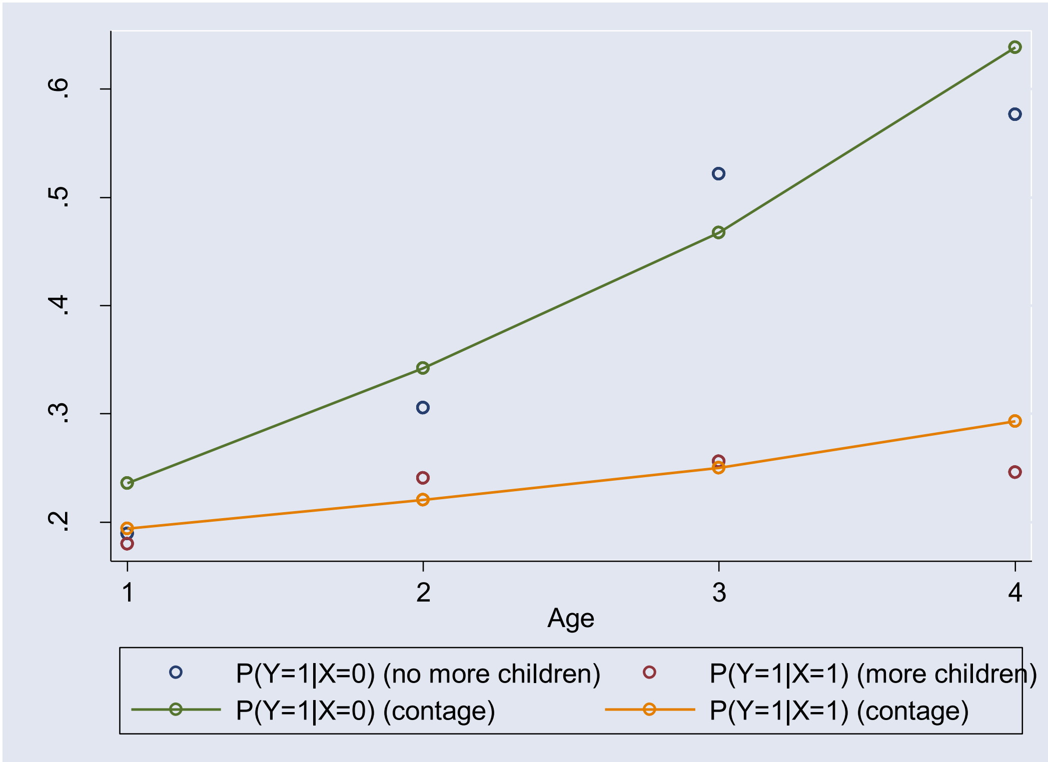
Two-factor model with interaction

```
.  xi: logit cuse contage i.more i.more*contage [freq=N],nolog
i.more                Imore_0-1    (naturally coded; Imore_0 omitted)
i.more*contage        ImXcon_#     (coded as above)
Note: Imore_1 dropped due to collinearity.
Note: contage dropped due to collinearity.

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(3)      =     136.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -933.57756                       Pseudo R2       =     0.0681

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 contage |   .0698143     .01144      6.103   0.000       .0473923    .0922362
 Imore_1 |   .7110262   .5082596      1.399   0.162      -.2851442    1.707197
ImXcon_1 |  -.0479913    .015438     -3.109   0.002      -.0782493   -.0177334
   _cons |  -2.573179   .4020974     -6.399   0.000      -3.361275   -1.785082

. est store M7

. lrtest M7 M6

likelihood-ratio test                                  LR chi2(1)  =      9.85
(Assumption: M6 nested in M7)                          Prob > chi2 =    0.0017
```

e) Is the interaction term significant?
f) What is the interpretation of the coefficient of the interaction term?

The two-factor model with interaction is shown graphically here (the points in the graph correspond to the predicted probabilities from the original model where age was treated as a categorical factor):

```
. predict phatcx
(option p assumed; Pr(cuse))
. gen phatcx0=phatcx if more==0
(16 missing values generated)
. gen phatcx1=phatcx if more==1
(16 missing values generated)
. label var phatcx1 "P(Y=1|X=1) (contage)"
. label var phatcx0 "P(Y=1|X=0) (contage)"
. sort age
. sc phatx0 phatx1 phatcx0 phatcx1 age , xlab() ylab() c(. . l l)
l1(Probability)
```



Now add a quadratic term for age to the model and then produce a graph with results from both models (with and without interaction):

```
. gen contage2=contage*contage

.  xi: logit cuse contage contage2 i.more i.more*contage [freq=N], nolog
i.more                Imore_0-1     (naturally coded; Imore_0 omitted)
i.more*contage        ImXcon_#      (coded as above)
Note: Imore_1 dropped due to collinearity.
Note: contage dropped due to collinearity.

Logit estimates                                   Number of obs   =       1607
                                                  LR chi2(4)      =     143.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -930.18024                       Pseudo R2       =     0.0715

------------------------------------------------------------------------------
    cuse |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 contage |   .2331551   .0651087      3.581   0.000       .1055445    .3607658
contage2 |  -.0024113   .0009398     -2.566   0.010      -.0042532   -.0005693
 Imore_1 |   1.292637   .5810191      2.225   0.026       .1538601    2.431413
ImXcon_1 |  -.0659373   .0176673     -3.732   0.000      -.1005645   -.0313101
   _cons |  -5.216035   1.123734     -4.642   0.000      -7.418513   -3.013557
------------------------------------------------------------------------------

. est store M8
. lrtest M7 M8
likelihood-ratio test                                  LR chi2(1)  =      6.79

(Assumption: M7 nested in M8)                          Prob > chi2 =    0.0091
```

g) Is the quadratic term significant?

Consider now the model where the interaction will encompass the quadratic term:

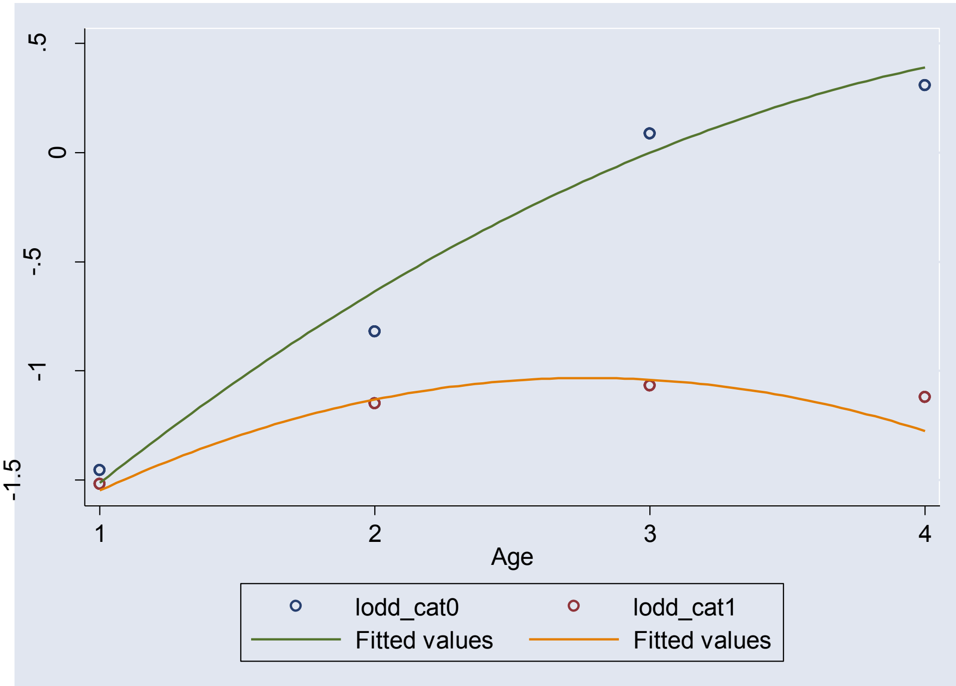
```
   . quietly xi: logit cuse contage contage2 i.more i.more*contage i.more*contage2
[freq=N]
   . est store M9

   . lrtest M8 M9

   likelihood-ratio test                                  LR chi2(1)  =      0.60

   (Assumption: M8 nested in M9)                          Prob > chi2 =    0.4399
```

h) Do you think that the inclusion of the quadratic interaction term in the model is required?

Stata code and graphs showing predicted log Odds by the last two models (`contage*more + contage^2` and `contage*more + contage^2*more`) along with predictions by the model with categorical age and its interaction with `more`:

```
qui xi: logit cuse i.more*i.age [freq=N],nolog
predict lodd_cat,xb
gen lodd_cat0=lodd_cat if more==0
gen lodd_cat1=lodd_cat if more==1

qui xi: logit cuse i.more*contage contage2 [freq=N],nolog
predict lodd_2cont,xb
gen lodd_2cont0=lodd_2cont if more==0
gen lodd_2cont1=lodd_2cont if more==1

qui xi: logit cuse i.more*contage i.more*contage2 [freq=N],nolog
predict lodd_3cont,xb
gen lodd_3cont0=lodd_3cont if more==0
gen lodd_3cont1=lodd_3cont if more==1
```

```
sc lodd_cat0 lodd_cat1 age , xlab() ylab() c(. . )  || qfit lodd_2cont0 age
|| qfit lodd_2cont1 age
```




```
sc lodd_cat0 lodd_cat1 age , xlab() ylab() c(. . )  || qfit lodd_3cont0 age
|| qfit lodd_3cont1 age
```