GENERALIZED LINEAR MODELS:Logistic Regression (2)

Γιώτα Τουλούμη

Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας

Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής

Ιατρική Σχολή Πανεπιστημίου Αθήνας

gtouloum@med.uoa.gr

Combining 2×2 contingency tables

Consider the following situation of contraceptive use according to desire for more children and age:

```
. sort age

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

-> age = <25

Contraceptive use (Yes/No)	Desires more children? No	Yes	Total
No	60	265	325
Yes	14	58	72
Total	74	323	397

-> age = 25-29

Contraceptive use (Yes/No)	Desires more children? No	Yes	Total
No	84	215	299
Yes	37	68	105
Total	121	283	404

-> age = 30-39

Contraceptive use (Yes/No)	Desires more children? No	Yes	Total
No	145	230	375
Yes	158	79	237
Total	303	309	612

-> age = 40-49

Contraceptive use (Yes/No)	Desires more children? No	Yes	Total
No	58	43	101
Yes	79	14	93
Total	137	57	194

The question that naturally arises is whether we should combine the information in those four tables and use all the available data in order to ascertain the effect of desire for more children on the probability of contraceptive use. However, if the association between contraceptive use and desire for more children is different in each age group, such an analysis would be inappropriate.

In general we have g tables (i=1,...,g) that are constructed as follows (g=4 in this example)

Disease	**Exposure** **Yes**	**No**	**Total**
Yes	a_i	b_i	N_{1i}
No	c_i	d_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

Estimates of the common odds ratio

Woolf (1995) proposed that a combined estimate of the log-relative odds be the weighted average of the observed odds ratio in each table, using weights inversely proportional to the estimated variances:

Weighted average with Weights the inverse of variance

$$\log Y = \frac{\sum w_i \log(\frac{a_i d_i}{b_i c_i})}{\sum w_i} \qquad where\, w_i = (\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i})^{-1}$$

The variance of this estimate is $var(\log Y) = (\sum w_i)^{-1}$

Mantel and Haenszel (M-H) proposed as a summary estimate the statistic:

$$Y = \frac{\sum a_i d_i / N_i}{\sum b_i c_i / N_i}$$

which can be recognized as a weighted average of the strata-specific odds with weights $b_i c_i / N_i$.

M-H estimate is not affected by zero entries and will give a consistent estimate even with large numbers of small strata. Its only major drawback is the lack of a robust variance estimate to accompany it.

Approximate confident intervals can be estimated using the Woolf method (logit scale):

$$\log(Y) \pm Z_{\alpha/2} / \sqrt{\sum w_i}$$

The test of homogeneity

We employ the following strategy:

1. Analyze the four tables separately. Based on the individual estimates of the odds ratios,

2. Test the hypothesis that the odds ratios in the four subgroups are sufficiently close to each other (they are *homogeneous*).

3a. If the assumption of homogeneity *is not rejected* then perform an overall (combined) analysis.

3b. If the homogeneity assumption *is rejected*, then perform separate analyses (the association of the two factors is different in each subgroup).

The test of homogeneity (continued)

The test of homogeneity is set-up as follows:

1. H_o: $OR_1=\ldots=OR_4$ (the four odds ratios do not have to be 1, just equal)

2. H_a: Any two odds ratios not equal (only two-sided alternatives are possible)

3. The test statistic $\chi^2=\sum_{i=1}^{g} w_i(y_i-\hat{y})^2$ has an approximate χ^2_{g-1}, where

a. $y_i=\ln(OR_i)=\ln\left[\frac{a_i d_i}{b_i c_i}\right]$ **b.** $w_i=\frac{1}{\left[\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}\right]}$ **c.** $\hat{y}=\frac{\sum_{i=1}^{g} w_i y_i}{\sum_{i=1}^{g} w_i}$

We use the individual odds ratios, producing a *weighted average*, weighing each of them inversely proportional to the square of their std. errors (one over their variance) to downweight odds ratios with high variability. High variability means low information.

4. Rejection rule. Reject the null hypothesis (conclude that the four subgroups are **not** *homogeneous* if $X^2>\chi^2_{g-1;0.05}$). **NOTE:** The X^2 is the usual one $x^2=\sum_i\{(y_i-\mu)^2/\sigma^2\}$

The Mantel-Haenszel test for association

The Mantel-Haenszel is based on the chi square distribution and the simple idea is that if there is no association between "exposure" and "disease", then the number of exposed individuals a_i contracting the disease should not be too different from $m_i = \frac{M_{1i} N_{1i}}{T_i}$ (expected number). The variance of m_i is

$$\sigma_i^2 = \frac{M_{1i} M_{2i} N_{1i} N_{2i}}{T_i^2 (T_i - 1)}.$$

Note: Conditional on fixed marginal totals the probability of the data consists a product of the non-central hypergeometrical distribution.

---

[1] If any of the cell counts is zero 3b becomes $w_i = \frac{1}{\left[\frac{1}{a_i + 0.5} + \frac{1}{b_i + 0.5} + \frac{1}{c_i + 0.5} + \frac{1}{d_i + 0.5}\right]}$

The Mantel-Haenszel test for association

The Mantel Haenszel test is constructed as follows:

1. H_o: OR=1
2. H_a: OR $\neq 1$ (only two-sided alternatives can be accommodated by the chi-square test)
3. The test statistic is $\chi^2_{MH} = \frac{\left[\sum_{i=1}^{g} a_i - \sum_{i=1}^{g} m_i\right]^2}{\sum_{i=1}^{g} \sigma_i^2}$,
4. **Rejection rule:** Reject H_o if $\chi^2_{MH} > \chi^2_{1;\alpha}$.

Example (continued): In the above example,

```
. 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

The relationship between contraceptive use and desire for more children is significant, but it is not constant (homogeneous) across levels of age (the test for homogeneity is significant). Notice that in this case, the test for homogeneity is significant, implying that the association of desire for more children and contraceptive use is not constant with age. We will see later how the presence of an interaction between age and desire for more children is handled by the logistic-regression analysis.

Logistic regression analysis

Analysis via logistic regression is a great deal more flexible. The fitted model is as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_o + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}_{\text{factors associated with age}} + \underbrace{\beta_4 X_4}_{\text{desire for more children}}$$

```
. 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

. lrtest, saving(3)
```

Interpretation of the model coefficients

The interpretation of the two-factor model is similar as that of the one factor.

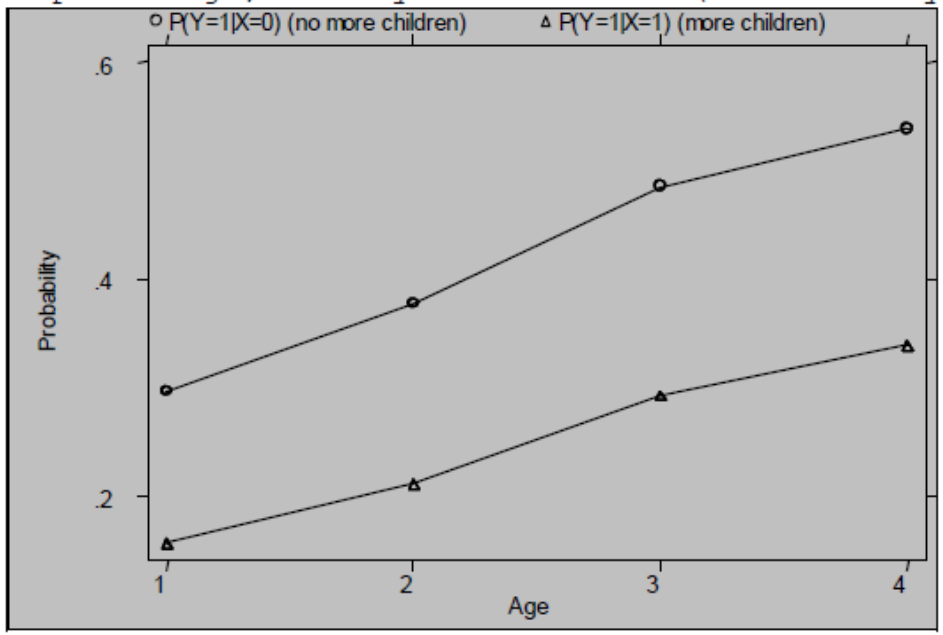
Both estimates of the slope of the line associated with each factor are interpreted as log-odds ratios. Thus, the odds ratio of using contraception versus not using, associated with the desire for more children versus desire for no more children is $e^{\hat{\beta}_4} = e^{-0.824} \approx 0.439$. That is, women that desire more children are less than half as likely to use contraception. This estimate is *adjusted* for age.

The estimates of each age group are themselves adjusted for all the other age groups as well as for desire for more children. They correspond to an odds ratio of the specific group compared to the reference group (women under 25 years of age). For example, adjusted for all the other factors (including the other age groups) women in age group 4 (40-49 years of age) are 2.78 ($=e^{1.023}$) i.e., almost three times as likely as women under the age of 25 to be using contraceptives.

Logistic regression analysis (continued)

The above model is shown graphically as follows:

```
. quietly xi: logit cuse i.age more

. 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 phat0 "P(Y=1|X=1) (more children)"
. graph phat0 phat1 age, xlab ylab border l1(Probability) c(ll)
```



Logistic regression analysis

Analysis via logistic regression is a great deal more flexible. The fitted model is as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_o + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}_{\text{factors associated with age}} + \underbrace{\beta_4 X_4}_{\text{desire for more children}}$$

```
. 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

. lrtest, saving(3)
```

1.44
2.22
2.77

Interpretation of the model coefficients

The interpretation of the two-factor model is similar as that of the one factor.

Both estimates of the slope of the line associated with each factor are interpreted as log-odds ratios. Thus, the odds ratio of using contraception versus not using, associated with the desire for more children versus desire for no more children is $e^{\hat{\beta}_4} = e^{-0.824} \approx 0.439$. That is, women that desire more children are less than half as likely to use contraception. This estimate is *adjusted* for age.

The estimates of each age group are themselves adjusted for all the other age groups as well as for desire for more children. They correspond to an odds ratio of the specific group compared to the reference group (women under 25 years of age). For example, adjusted for all the other factors (including the other age groups) women in age group 4 (40-49 years of age) are 2.78 ($= e^{1.023}$) i.e., almost three times as likely as women under the age of 25 to be using contraceptives.

The two-factor model with interaction

We introduce one more addition to the model above (X_1-X_3 are associated with age and X_4 is the desire for more children). Thus,

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_o+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_1X_4+\beta_6X_2X_4+\beta_7X_3X_4$$
$$=\begin{cases}\beta_o+(\beta_1+\beta_5)X_1+(\beta_2+\beta_6)X_2+(\beta_3+\beta_7)X_3+\beta_4 & \text{if } X_4=1\\ \beta_o+\beta_1X_1+\beta_2X_2+\beta_3X_3 & \text{if } X_4=0\end{cases}$$

```
. 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)
i.age*i.more         IaXm_#-#      (coded as above)



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
------------------------------------------------------------------------------
.lrtest, saving(4)
```

1.89
4.64
6.84
0.94

Interpretation of model coefficients

Interpretation of the interaction coefficients is not straightforward even in a relatively simple model as the one considered here. Consider the overall model:

$$\text{logit}(\pi) = \beta_o + \beta_1 age\_2 + \beta_2 age\_3 + \beta_3 age\_4 + \beta_4 more + \beta_5 age\_2*more + \beta_6 age\_3*more + \beta_7 age\_4*more .$$

The model for women that desire more children (X_4=1) becomes:

$$\text{logit}(\pi) = (\beta_o + \beta_4) + (\beta_1 + \beta_5) age\_2 + (\beta_2 + \beta_6) age\_3 + (\beta_3 + \beta_7) age\_4$$

whereas for women that do not desire more children (X_4=0), it becomes:

$$\text{logit}(\pi) = \beta_o + \beta_1 age\_2 + \beta_2 age\_3 + \beta_3 age\_4$$

For women aged <25 years, the OR of using contraceptives for women desiring more children vs those that do not is $e^{\hat{\beta}_4} = e^{-0.064} = 0.94$. That is, women who desire more children are 6.2% less likely than women who do not to use contraceptives.

Interpretation of model coefficients (2)

The corresponding OR for women aged 25-29 years is:

$$e^{\hat{\beta}_4+\hat{\beta}_5} = e^{-0.064-0.267} = e^{-0.064} * e^{-0.267} = e^{=-0.33} = 0.718$$

therefore, women who desire more children are 28.20% less likely than women who do not to use contraceptives, among those aged 25-29 years. The effect of the interaction term is multiplicative.

Also, the effect of the interaction in the group of women 25-29 years of age is multiplicative on the main odds ratio by a factor of $e^{\hat{\beta}_5} = e^{-0.267} \approx 0.766$. This can also be derived by comparing the odds ratios of these two groups versus the reference group (<25 years old). These are 1.887 ($= e^{\hat{\beta}_1} = e^{0.635}$) among women that want no more children and 1.445 ($= e^{\hat{\beta}_1+\hat{\beta}_5} = e^{(0.635-0.267)}$) among women that desire more children. The ratio of these two odds ratios is 76.6%.

Presentation of results from models with interactions

Age (years)	No More	Yes More
<25	1	$e^{\beta_4} = e^{-0.64} = 0.94$
25-29	$e^{\beta 1} = e^{0.64} = 1.89$	$e^{\beta_1} * e^{\beta 4} * e^{\beta_5} = 0.94 * 1.89 * 0.77 = 1.35$
30-39	$e^{\beta_2} = e^{1.54} = 4.67$	$e^{\beta_2} * e^{\beta_4} * e^{\beta_6} = 0.94 * 4.67 * 0.34 = 1.47$
40-49	$e^{\beta_3} = e^{1.76} = 5.84$	$e^{\beta_3} * e^{\beta_4} * e^{\beta_7} = 0.94 * 5.84 * 0.26 = 1.39$

Therefore, OR for Yes/No:

Age (years)	OR
<25	$e^{\beta_4} = e^{-0.64} = 0.94$
25-29	$e^{\beta_4} * e^{\beta_5} = 0.94 * 0.77 = 0.75$
30-39	$e^{\beta_4} * e^{\beta_6} = 0.94 * 0.34 = 0.32$
40-49	$e^{\beta_4} * e^{\beta_7} = 0.94 * 0.26 = 0.24$

OR for each age group comparing to <25 for women who desire more children:

25-29/<25	$e^{\beta_1} * e^{\beta_5} = 1.89 * 0.77 = 1.43$
30-39/<25	$e^{\beta_2} * e^{\beta_6} = 4.67 * 0.34 = 1.57$
40-49/<25	$e^{\beta_3} * e^{\beta_7} = 5.84 * 0.26 = 1.49$

Model selection

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

1. Model with `more` versus the null model

```
. lrtest, model(0) using(1)
Logit:  likelihood-ratio test                     chi2(1)     =      91.67
                                                  Prob > chi2 =     0.0000
```

2. Model with `age` versus the null model

```
. lrtest, model(0) using(2)
Logit:  likelihood-ratio test                     chi2(3)     =      79.19
                                                  Prob > chi2 =     0.0000
```

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

```
. lrtest, model(1) using(3)
Logit:  likelihood-ratio test                     chi2(3)     =      37.21
                                                  Prob > chi2 =     0.0000
```

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

```
. lrtest, model(2) using(3)
Logit:  likelihood-ratio test                     chi2(1)     =      49.69
                                                  Prob > chi2 =     0.0000
```

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

```
. lrtest, model(3) using(4)
Logit:  likelihood-ratio test                     chi2(3)     =      16.79
                                                  Prob > chi2 =     0.0008
```

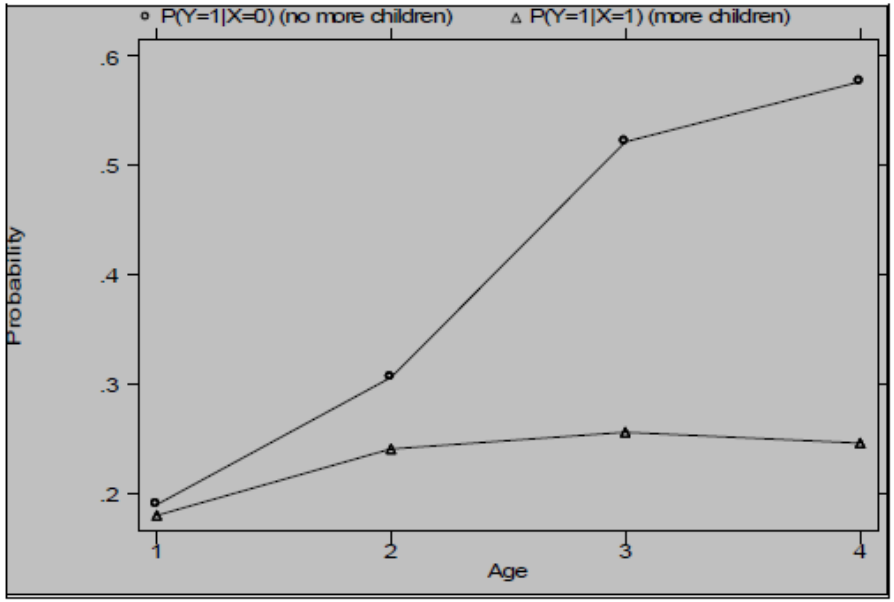
Model selection

Model	Likelihood-ratio	Difference	Df	p-value
Two factors (with interaction)	145.67	—		—
Two factors (no interaction)	128.88	16.82	3	0.0008
Age	79.19	66.48	3	0.0000
Desires more children?	91.67	54.00	1	0.0000

From the table above we conclude that there is significant interaction between age and desire for more children. Note that the same results would be produced by considering *deviances* of these models (using the `glm` command in STATA)

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

```
. predict phatx
(option mu assumed; predicted mean 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)"
. graph phatx0 phatx1 age , border xlab ylab c(ll) l1(Probability)
```



This means that while the likelihood of contraceptive use increases with age, it does so more quickly among women that want no more children. This is why the no-interaction model was inadequate.

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. We do not have access to the individual ages in this data, so we follow the approach in Rodriguez (2000) and substitute for age the interval means as follows:

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

The models that we consider are (Z is the covariate age and X the factor desire for more children)

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_o+\beta_1 X\,,\ \log\left(\frac{\pi}{1-\pi}\right)=\beta_o+\beta_1 Z$$

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_o+\beta_1 X+\beta_2 Z$$

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_o+\beta_1 X+\beta_2 Z+\beta_3 XZ=\begin{cases}(\beta_o+\beta_1)+(\beta_2+\beta_3)Z & \text{if } X=1\\ \beta_o+\beta_2 Z & \text{if } X=0\end{cases}$$

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     8.54   0.000     .0467486    .0745934
       _cons |  -2.672667   .2332492   -11.46   0.000    -3.129827   -2.215507
------------------------------------------------------------------------------
.lrtest, saving(5)
```

Notice that now the covariate age is associated with one degree of freedom. This is a much more parsimonious description of the relationship and, if the relationship is linear, results in more powerful tests.

Interpretation of estimates in a continuous-factor model

The estimate of the slope associated with age is $\hat{\beta}_1 \approx 0.061$. This is interpreted as the log odds ratio for each year of age. The fitted model is $\log\left[\frac{\pi}{1-\pi}\right] = \beta_o + \beta_1 Z$. Thus, two women with difference of one year in age will have odds ratio for use of contraceptive methods (Y) as follows:

$$\hat{\psi} = \frac{\hat{\pi}(Z+1)/1-\hat{\pi}(Z+1)}{\hat{\pi}(Z)/1-\hat{\pi}(Z)} = \frac{\dfrac{e^{\hat{\beta}_o+\hat{\beta}_1(Z+1)}}{1+e^{\hat{\beta}_o+\hat{\beta}_1(Z+1)}} \Big/ \dfrac{1}{1+e^{\hat{\beta}_o+\hat{\beta}_1(Z+1)}}}{\dfrac{e^{\hat{\beta}_o}}{1+e^{\hat{\beta}_o+\hat{\beta}_1 Z}} \Big/ \dfrac{1}{1+e^{\hat{\beta}_o+\hat{\beta}_1 Z}}} = \frac{e^{\hat{\beta}_o+\hat{\beta}_1}}{e^{\hat{\beta}_o}} = e^{\hat{\beta}_1}$$

where $\hat{\pi}(z) = P(Y=1 \mid Z=z)$.

Thus, the odds ratio for each year of age is $e^{0.061} = 1.063$. Interpreted as an approximate relative risk this means that each year, women are on average 6.3% *more* likely to use contraception. If we wanted to calculate the odds ratios for 10 years, that would be $e^{(10)0.061} = 1.84$. That is, for every decade of life, women are 84% *more* likely to use contraception.

Two-factor model with no interaction

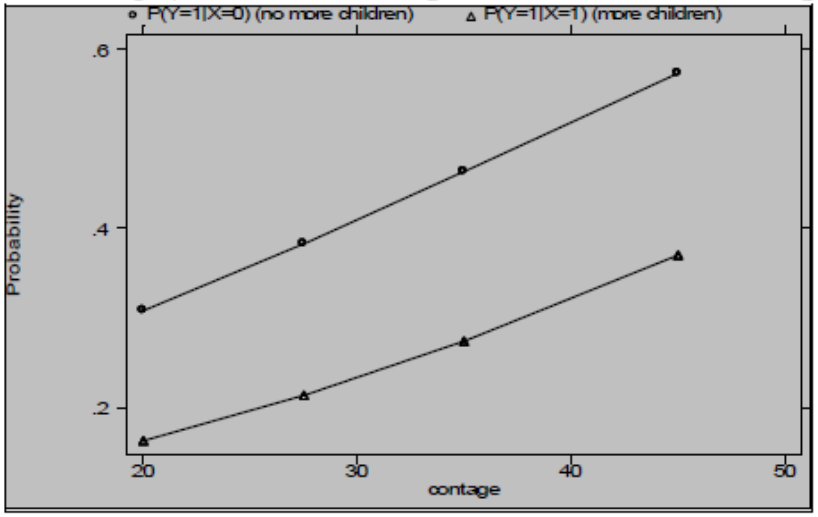
The model including both age and desire for more children is given as follows:

```
.  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    -7.05   0.000    -1.055429   -.5963665
     contage |   .0441062    .007529     5.86   0.000     .0293497    .0588627
       _cons |  -1.690756   .2701814    -6.26   0.000    -2.220302    -1.16121
------------------------------------------------------------------------------.
 lrtest, saving(6)

. lrtest, model(5) using(6)
Logit:  likelihood-ratio test                     chi2(1)     =      49.90
                                                  Prob > chi2 =     0.0000
```

The increase in the likelihood-ratio statistic is 126.69-76.79=49.90 which is extremely significant based on a chi-square distribution with one degree of freedom.

The model assumes two parallel lines for the relationship of the probability of contraceptive use by age in the two groups (women wanting more children versus women wanting no more children).

```
. 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=0)  (no more children)"
. label var yhat0 "P(Y=1|X=0)  (no more children)"
. sort more age
. graph yhat0 yhat1 contage, c(ll) xlab ylab l1("Probability") border
```



The probability of using contraceptives increases with age. It is lower in all ages among women that want more children by a constant amount (β_1).

Two-factor model with interaction

A two-factor model that does not assume parallelism between the two groups of women (those wanting more children versus those that do not) introduces an interaction term as follows:

```
.  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

. lrtest, saving(7)

. lrtest, model(6) using(7)
Logit:  likelihood-ratio test                     chi2(1)     =       9.85
                                                  Prob > chi2 =     0.0017
```

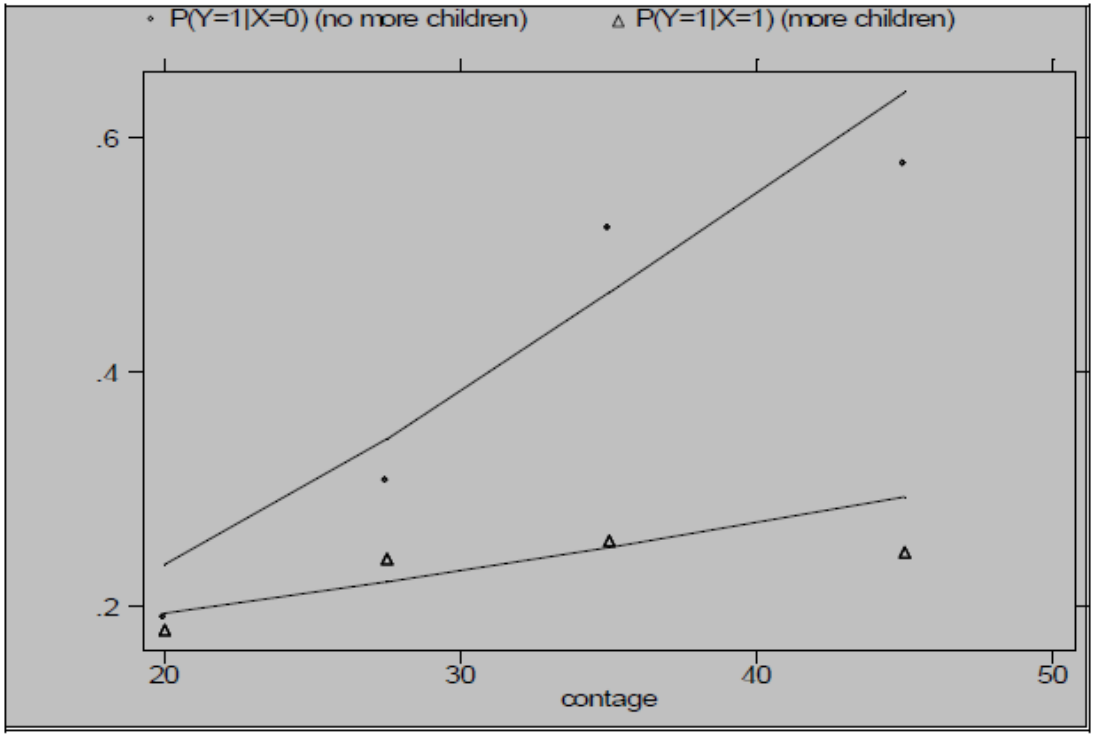
Interpretation of the model coefficients

1. The likelihood-ratio for this model is 136.54. This is an increase of 136.54-126.69=9.85 for one additional degree of freedom. This compared to a chi-square distribution with one degree of freedom is associate with a p-value of 0.0017, which is highly significant. The interaction model is a significant improvement over the parallel lines (no-interaction) model.
2. From the graph above we see what the effect of interaction is: the coefficient $\hat{\beta}_3 = -0.048$ decreases the slope of the line corresponding to the group that desires more children (X=1). The slope in this group is $\hat{\beta}_1 + \hat{\beta}_3 = 0.0698 - 0.048 = 0.0218$, while in the other group (X=0) is $\hat{\beta}_1 = 0.0698$. Thus, the increase in the probability of using contraceptive is steeper with increasing age among women that desire no more children.

Two-factor model with interaction (continued)

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):

```
. graph phat0 phat1 phatx0 phatx1 contage, c(..ll) s(oTii) xlab ylab
> border l1(probability)
```



Higher-order models

We can add a quadratic term for age to the model as follows:

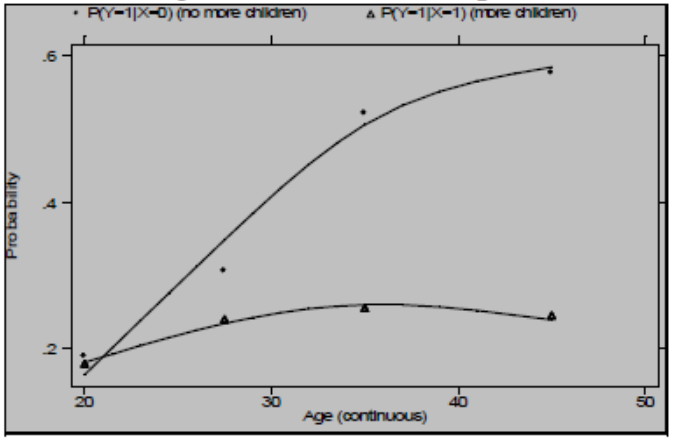
$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_o+\beta_1 X+\beta_2 Z+\beta_3 XZ+\beta_4 Z^2=\begin{cases}(\beta_o+\beta_1)+(\beta_2+\beta_3)Z+\beta_4 Z^2 \text{ if } X=1\\ \beta_o+\beta_2 Z+\beta_4 Z^2 \text{ if } X=0\end{cases}$$

```
. 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
------------------------------------------------------------------------------
. lrtest, saving(8)
. lrtest, model(7) using(8)
Logit:  likelihood-ratio test                     chi2(1)     =       6.79
                                                  Prob > chi2 =     0.0091
```

More models

A model where the interaction will encompass the quadratic term is done as follows:

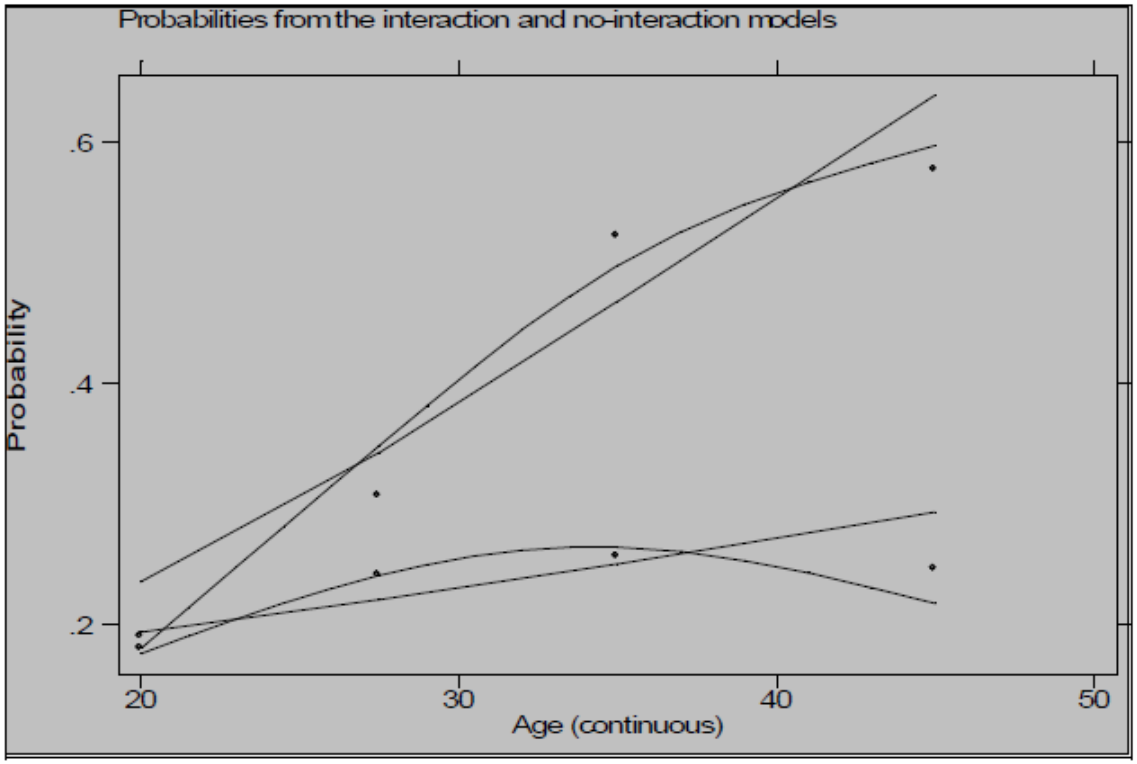
```
. quietly xi: logit cuse contage contage2 i.more i.more*contage i.more*contage2 [freq=N]
>  [freq=N]
. lrtest, saving(9)
. lrtest, model(8) using(9)
Logit:  likelihood-ratio test                         chi2(1)     =       0.60
                                                      Prob > chi2 =     0.4399

. predict phat3
(option p assumed; Pr(cuse))
. gen phat30=phat3 if more==0
(16 real changes made)
. replace phat31=phat3 if more==1
(16 real changes made)
. sort age cuse more
. graph phat0 phat1 phat30 phat31 contage, c(..ss) xlab ylab border s(oTii) l1(Probability)
```



This model has a likelihood ratio of 143.93, only 0.60 larger than before (p-value 0.4399). The previous quadratic model seems to fit the data adequately.

Interpretation of the quadratic model

```
.  graph phat0 phat1 phatx20 phatx21 phatx0 phatx1  contage, c(..ssll) s(ooiiii)
>  xlab ylab border l1("Probability") gap(4)  t1("Fitted curves generated from the
>  interaction and no-interaction models")
```



The likelihood ratio increase for the quadratic term is 143.33-136.54=6.79, which is associated with a tail of a chi-square with one degree of freedom with p value of 0.0091. Thus, the addition of the quadratic term is a significant improvement in the model.

Prospective versus retrospective studies

Prospective study: an exposed group of subjects is selected TOGETHER with a comparable group of non-exposed individuals. The progress of each group is monitored, often over a prolonged period, with a view towards comparing the incidence of disease in the two groups.

In this way **the row totals, giving the number of subjects in each category of exposure, are fixed by design.** The column totals are RANDOM, reflecting the incidence of disease in the overall population, weighted according to the sizes of exposure groups in the sample.

Retrospective study: Disease & disease-free individuals are selected-often from hospital records over a period of time. In this design, **the column totals are fixed by design and the row totals are random**, reflecting the frequency of exposure in the population, weighted according to the sizes of the disease groups in the sample.

Adaptation of the logistic model to case-control studies

In a cohort (prospective study) the logistic model for a specified regression vector **x** will be of the form:

$$P(D/\underset{\sim}{x}) = \frac{e^{\alpha+\beta^T \cdot \underset{\sim}{x}}}{1+e^{\alpha+\beta^T \cdot \underset{\sim}{x}}}$$

for the probability of contracting the disease given covariates $\underset{\sim}{x}$.

The above model is specified for data sampled prospectively. Suppose now, the data are sampled retrospectively. Let

$$Z = \begin{cases} 1 & \text{if individual is sampled} \\ 0 & \text{otherwise} \end{cases}$$

then

$$P(Z=1 \mid D) = \pi_0, \qquad P(Z=1 \mid \overline{D}) = \pi_1$$

(it is essential here that the sampling proportions depend on D and not on x.).

Adaptation of the logistic model to case-control studies (continue)

Use Bayes's theorem now to find the disease frequency among sampled individuals who have a specified covariate vector x.

$$P(D|Z) = \frac{P(DZ)}{P(Z)} = \frac{P(Z|D)P(D)}{P(Z|D)P(D) + P(Z|\bar{D})P(\bar{D})}$$

$$P(D \mid Z=1) = \frac{P(Z=1|D,x)P(D|\underline{x})}{P(Z=1|D,x)P(D|x) + P(Z=1|\bar{D},x)P(\bar{D}|x)}$$

Devide numerator and denominator by π_1

$$= \frac{\pi_0 \exp(\alpha + \beta^T x)}{\pi_1 + \pi_0 \exp(\alpha + \beta^T x)}$$

$\exp(\ln(\frac{\pi_o}{\pi_1}))$

$$= \frac{\exp(\alpha^* + \beta^T x)}{1 + \exp(\alpha^* + \beta^T x)}$$

where $\alpha^* = \alpha + \log\left(\frac{\pi_0}{\pi_1}\right)$.

In other words, the disease probabilities for those in the sample continue to be given by the logistic model with precisely the same βs, albeit a different value for α.

Efficient design

For **rare diseases** the retrospective design is substantially more efficient than the prospective design. This is because the investigator has access via hospital records to all cases of the disease recorded over a substantial period of time.

In the case of rare diseases, it is common to take a 100% sample of the diseased individuals and to compare these with a similar sized sample of disease – free subjects.

For a prospective design to be effective, a large number of initially healthy subjects must be followed for a prolonged period in order that a sufficiently large number of subjects may eventually fall victim to the disease.

REMINDER: For rare diseases the odds ratio is an approximation of the risk ratio