Answers to questions - Laboratory session 6

Analysis using a 2×2 table or logistic regression

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

```
. di chi2tail(1,92.6442)
6.259e-22
```

b) Compare the chi-square statistic in the `logit` command output with the one given in 2×2 table analysis.

The two statistics have almost the same value.

c) Calculate the Odds for the use of contraceptives in the two "`more`" categories.

```
. *more=0
. di exp(-.1863643)
.82997118

. *more=1
. di exp(-.1863643-1.048629)
.29083671
```

d) Calculate the Odds Ratio. Now use the 2×2 table data to produce the Odds Ratio. Compare the two OR's.

Using previous results

```
. di exp(-.1863643-1.048629)/exp(-.1863643)
.35041784
```

Or using only the "`more`" coefficient

```
. di exp(-1.048629)
.35041784
```

Now using the 2×2 table

```
. di (347*219)/(288*753)
.35041777
```

The two approaches gave the same result.

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?

We can check the z-statistic value and the associated p-value. The z-statistic equals the

coefficient divided by its standard error.

```
. di -1.048629/.110672
-9.4751066
```

The asymptotic distribution of this statistic is standard normal

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

By exponentiating the 95% C.I. for the "`more`" coefficient in the logit command output.

```
. di exp(-1.265542)
.28208636
. di exp(    -.831716)
.43530167
```

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

logOdds(use of contraceptives) = $\beta_0 \Rightarrow$ Odds(use of contraceptives) = exp(β_0) = .46090907

Using the 2×2 table data

```
. tab cuse [freq=N]

Contracepti |
     ve use |
   (Yes/No) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |       1100       68.45       68.45
        Yes |        507       31.55      100.00
------------+-----------------------------------
      Total |       1607      100.00

. di 507/1100
.46090909
```

h) Calculate the $-2\log\lambda$ 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".

$-2\log\lambda$=-2*(-1001.8468-(-956.00957))= 91.67446 ≅ $(-9.475)^2$

Analysis using a 2×c table or logistic regression

a) What is the value of the likelihood ratio statistic? Compare it to the appropriate distribution in order to obtain the relevant p-value.

LR=79.19 . Assymptotical distribution chi-square with 3 degrees of freedom.

```
. di chi2tail(3,79.19)
4.579e-17
```

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.

Using the logit coefficients …

```
. di exp(.4606758)
1.5851449

. di exp(1.048293)
2.8527773

. di exp(1.424638)
4.156353
```

Using cross products in the 2×4 table …

```
. di (325*105)/(299*72)
1.5851449

. di (325*237)/(375*72)
2.8527778

. di (325*93)/(101*72)
4.1563531
```

The results are exactly the same.

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 check the significance of each group by using the results (z-statistics and relevant p-values) of the individual Wald tests for the coefficients, given in the `logit` command output

Two factors

a) Is the relationship between contraceptive use and desire for more children significant?

According to the M-H analysis the relationship between contraceptive use and desire for more children is significant (M-H chi-square=50.36, p-value<0.000). At the same time the test of homogeneity is significant (the OR is not constant across the age levels) so the M-H analysis is inappropriate for this case.

b) The test for homogeneity is significant. What is the interpretation of this result?

The relationship between contraceptive use and desire for more children is not constant across age levels.

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)

```
quietly xi: logit cuse i.age more [freq=N]
predict xphat,xb
generate xphat0=xphat if more==0
generate xphat1=xphat if more==1
label var xphat0 "logOdds(Y=1|X=0)"
label var xphat1 "logOdds(Y=1|X=1)"
sort age
sc xphat0 xphat1 age, xlab() ylab()  l1(log Odds) c(l l)
```

d) Calculate the adjusted for age estimate of the odds ratio of using contraception versus not using, associated with the desire for more children versus desire for no more children.

```
. di exp( -.824092)
.43863309
```

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.

```
. di exp(1.022618)
2.7804645
```

f) What is the underlying assumption of the previous model about the difference between the two “more” groups acrross the four age group categories.

Since we do not include an interaction term we assume that the difference between the two “more” groups acrross the four age group categories remains constant (in the logit scale).

The two-factor model with interaction

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.

i. $OR_{(40\text{-}49/<25 \mid more=1)} = \exp(1.764292 \quad -1.367148) = 1.49$
ii. $OR_{(40\text{-}49/<25 \mid more=0)} = \exp(1.764292 \quad) = 5.83$
The interaction term (IaXm_4_1) coefficient if exponentiated equals to the ratio of the two OR's above.

b) What is the main difference between the models with and without the interaction term?

In the model with the interaction term included we allow for changes in the magnitude of difference between the two more categories acrross the age groups.

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

```
xi: logit cuse i.age i.more i.age*i.more  [freq=N],nolog
predict zphatx,xb
gen zphatx0=exp(zphatx) if more==0
gen zphatx1=exp(zphatx) if more==1
label var zphatx1 "Odds(Y=1|X=1)"
label var zphatx0 "Odds(Y=1|X=0)"
sort age
sc zphatx0 zphatx1 age , xlab() ylab() c(l l) l1(Odds)
```

Model selection

a) Fill the following table. Pn+1 is the "smaller" model which is nested in the previous model Pn 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)	-929,01	16,7888	3	0,00078
Two factors (no interaction)	-937,4	37,21016	3	4,1536E-08
Desires more children?	-956,01	91,67446	1	1,0218E-21
Null model	-1001,8			

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

According to the likelihood ratio test p-value=0.00078<0.05, so the interaction term is significant.

Analysis of covariance-type models

a) What is the interpretation of the "contage" coefficient?

This coefficient gives us the log (Odds Ratio) for 1 year change in age.

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?

The covariate age (continuous) 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.
The main difference is that using age as a continuous variable we assume that the rate of change in the log(Odds) scale is constant for the entire range of ages in our dataset.

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.

The z-statistic in the `logit` command output equals –7.052 and the associated p-value is less than 0.001. Thus the "`more`" variable is highly significant.
The z-statistic in the `logit` command output equals the square root of the chi-square statistic in the `lrtest` command output.

d) Why are the lines not exactly straight?

Because we are plotting probabilities instead of log (odds).

e) Is the interaction term significant?

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 associated 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.

f) What is the interpretation of the coefficient of the interaction term?

From the graph 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).

g) The slope in this group is $\hat{\beta}_1 + \hat{\beta}_3 = 0.0698 - 0.0480 = 0.0218$, while in the other group (X=0) is $\hat{\beta} = 0.0698$.
Thus, the increase in the probability of using contraceptive is steeper with increasing age among women that desire no more children.

h) Is the quadratic term significant?

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.

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

This model has a likelihood ratio of 143.93, only 0.60 larger than before (p-value 0.4399). Thus the quadratic interaction term is not required in the model.