Biost 536: Categorical Data Analysis in Epidemiology
Emerson, Fall 2014

Homework #4 Key
November 4, 2014

Written problems: To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 11:30 pm on Sunday, November 9, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

> *On this (as all homeworks) Stata / R code and unedited Stata / R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*
>
> ***In all problems requesting "statistical analyses" (either descriptive or inferential), you should present both***
>
> - ***Methods:** A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.*
> - ***Inference:** A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details.*

Questions refer to analyses of the data in the file `infarcts.txt` that is located on the class webpages. We are interested in associations between prevalence of infarct-like lesions on MRI and various predictors. For this homework, we will presume that any missing data is missing completely at random (MCAR) in this dataset and hence ignorable.

***Instructions for grading:** On this key I place the total points to be awarded for a particular problem just prior to the answer for the problem. I also sometimes provide very specific criteria for awarding points. Remember the purpose of peer grading is to have the grader focus more closely on answers to the problems that were potentially different from those that he/she provided, and to identify areas where the answers on the paper being graded might not be correct. It is not of value to anyone to be overly "lenient" in the grading. So while it is not appropriate to capriciously deduct points, it is equally not appropriate to give points only because "Well, they tried" (unless, of course, such a criterion is specified in the grading instructions). We welcome students' questions regarding the appropriateness of answers during the grading process. And grades assigned by the peer grader can always be appealed.*

In providing answers below, the required answers are in* bold face type*. Further verbage that is in* regular italics type ***is additional information that you are in fact responsible for knowing on exams, but was not really required for the homework assignment. That additional information may help you decide whether the answers you are grading are valid, however.

In each problem requiring both description of methods and results, full credit should only be given when both aspects are appropriately addressed. The grader should be able to determine the exact analysis method from their description.

Within each problem, I provide Stata code that I used to produce the answers. This is for your later reference when trying to understand Stata. I did not expect (and in fact do not tolerate) the Stata

output as solutions to the homework assignment. It should not have been included in a homework set turned in by a student.

1. Fit a logistic regression model investigating prevalence of infarcts as a function of age (modeled continuously) and coronary heart disease (modeled as dummy variables). Provide a scientific interpretation of each of the regression coefficients, including a description of the intercept in the model. (You do not need to describe the methods, or provide CI or p values.)

***Instructions for grading*: This problem is worth 20 points. Issues to consider:**

- ***I provide answers as if the dummy variables had been fit using chd=0 as the reference group (so reference group is no prior history of angina or MI). I am guessing that no student used any other group as the reference group (and, indeed, I believe it definitely advantageous not to). But if they did, I would expect the grader to ascertain that their answers corresponded to that alternative approach.***
- ***Similarly, I am guessing that no students would have recoded age, but I do provide interpretations as if that had been done. I do expect all students to understand how the re-parameterization of the model affects the interpretations, so be sure you read and understand those additional parts.***
- ***I provide interpretation based on odds and odd ratios, rather than log odds and log odds ratios. I believe this is greatly to be preferred.***

***Stata code and output*:**

```
. logistic infarcts age i.chd
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(3)      =      76.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -2355.9742                       Pseudo R2       =     0.0160
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.04562   .0070283     6.64   0.000     1.031935    1.059486
             |
         chd |
          1  |   1.284301   .1441801     2.23   0.026     1.030642    1.600389
          2  |   1.768299   .2012159     5.01   0.000     1.414806    2.210113
------------------------------------------------------------------------------

. logit infarcts age i.chd
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(3)      =      76.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -2355.9742                       Pseudo R2       =     0.0160
------------------------------------------------------------------------------
    infarcts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |     .04461   .0067217     6.64   0.000     .0314358    .0577843
             |
         chd |
          1  |   .2502144   .1122635     2.23   0.026      .030182    .4702468
          2  |   .5700182   .1137906     5.01   0.000     .3469926    .7930437
             |
       _cons |  -4.151513   .5073745    -8.18   0.000    -5.145948   -3.157077
------------------------------------------------------------------------------

. di exp(-4.151513)
.01574058
```

Ans: **Interpretations of the various regression parameters are as follows:**

- **(intercept) In extrapolating from the data on 65 – 100 year olds, the odds of prevalent infarcts in newborns without prior history of angina or MI is estimated to be 0.0157.** *(This*

corresponds to an estimated probability of 1.55%, however it is so far outside the range of our data as to be totally irrelevant.)

- **(slope for age) When comparing groups of two different ages but having similar prior history of coronary heart disease, we estimate that the odds of prevalent infarcts is 4.56% higher (OR = 1.0456) for every year that one group is older than the other.**
- **(slope for chd=1) When comparing a group of patients of a specific age who have a prior history of angina (but without prior history of MI) to a group of patients of that same age who have no prior history of either angina or MI, we estimate that the odds of prevalent infarcts is 28.4% higher (OR = 1.284) in the group with prior history of angina.**
- **(slope for chd=2) When comparing a group of patients of a specific age who have a prior history of MI (with or without prior history of angina) to a group of patients of that same age who have no prior history of either angina or MI, we estimate that the odds of prevalent infarcts is 76.8% higher (OR = 1.768) in the group with prior history of MI.**

Note that I could have fit a model that would make the intercept more interpretable by creating a variable that measures age as "years above 65". Because age would still be measured in years, this will not change the slope for the continuously modeled linear term in age. It would change the intercept, however, to make it refer to a group of patients who were 65 years old without prior history of angina or MI. I demonstrate this using the log odds scale as returned by Stata's "logit" command.

```
. g yrsgt65= age - 65
. logit infarcts yrsgt65 i.chd
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(3)      =      76.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -2355.9742                       Pseudo R2       =     0.0160

------------------------------------------------------------------------------
    infarcts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yrsgt65 |     .04461   .0067217     6.64   0.000     .0314358    .0577843
             |
         chd |
          1  |    .2502144   .1122635     2.23   0.026      .030182    .4702468
          2  |    .5700182   .1137906     5.01   0.000     .3469926    .7930437
             |
       _cons |   -1.25186   .0794762   -15.75   0.000    -1.407631    -1.09609
------------------------------------------------------------------------------

. di exp(-1.25186)
.28597239
```

- *(intercept) The odds of prevalent infarcts in 65 year olds without prior history of angina or MI is estimated to be 0.286. (This corresponds to an estimated probability of 22.2%, Note that this estimate corresponds exactly to the fitted value for 65 year olds without prior CHD as estimated in the first model I gave: -4.151513 + 65 * 0.04461 = -1.251863.."Re-centering" age is just a re-parameterization of the exact same model.*
- *(all slopes stayed the exact same.***)**

Alternatively, I could have fit a model that would make the intercept more interpretable by creating a variable that measures age as "decades above 65". This will change the slope for the continuously modeled linear term in age, but the intercept would still be as given immediately above (0 years greater than 65 implies 0 decades greater than 65). I demonstrate the effect of this "re-scaling" of age on the odds scale as returned by Stata's "logistic" command.

```
. g decgt65 = yrsgt65 / 10
. logistic infarcts decgt65 i.chd
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(3)      =      76.45
                                                  Prob > chi2     =     0.0000
```

```
Log likelihood = -2355.9742                     Pseudo R2       =     0.0160
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     decgt65 |   1.562208   .1050067     6.64   0.000      1.36938     1.78219
             |
         chd |
           1 |   1.284301   .1441801     2.23   0.026     1.030642    1.600389
           2 |   1.768299   .2012159     5.01   0.000     1.414806    2.210113
```

- *(intercept stays the same, but you would have to look at "logit" output to see it)*
- *(slope for age) When comparing groups of two different ages but having similar prior history of coronary heart disease, we estimate that the odds of prevalent infarcts is 56.2% higher (OR = 1.562) for every decade that one group is older than the other.(This is exactly what we would have obtained from a 10 year difference in age from the first model that measured age in years: $1.0456^{10} = 1.562$ or exp(10 * 0.04461) = 1.562) Re-scaling and re-centering age is just a reparameterization of the exact same model, and all fitted values will remain the same.*
- *(all slopes related to CHD stayed the exact same.)*

Lastly, I could have fit a model that used the group with prior history of MI as the "reference" group. This would dummy variables for the no-CHD group and the angina only group. I demonstrate the effect of this change of dummy variables on the odds scale as returned by Stata's "logistic" command.

```
. logistic infarcts decgt65 ib2.chd
Logistic regression                             Number of obs   =       3775
                                                LR chi2(3)      =      76.45
                                                Prob > chi2     =     0.0000
Log likelihood = -2355.9742                     Pseudo R2       =     0.0160
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     decgt65 |   1.562208   .1050067     6.64   0.000      1.36938     1.78219
             |
         chd |
           0 |   .5655152   .0643503    -5.01   0.000     .4524655    .7068105
           1 |   .7262916   .1085886    -2.14   0.032       .54181    .9735875
```

- *(intercept stays the same, but you would have to look at "logit" output to see it)*
- *(slope for age stays the same because I have not modeled any interactions)*
- *(slope for chd=0) When comparing a group of patients of a specific age who have no prior history of angina or MI to a group of patients of that same age who have a prior history of MI, we estimate that the odds of prevalent infarcts is 43.4% lower (OR = 0.566) higher in the group without prior history of angina or MI. (Note that this is the same relationship as was estimated in the first parameterization, because 1 / 0.5655152 = 1.768, which was the slope for the chd=2 group in the first model).*
- *(slope for chd=1) When comparing a group of patients of a specific age who have a prior history of angina but no prior history of MI to a group of patients of that same age who have prior history of MI, we estimate that the odds of prevalent infarcts is 27.4% lower (OR = 0.726) in the group with only prior history of angina (Note that this is truly the same relationship as was estimated in the first parameterization, but we have to work a little harder to show this: In the first parameterization we can divide the OR comparing group 1 to group 0 by the OR comparing group 2 to group 0 to obtain the OR comparing group 1 to group 2: 1.284301 / 1.768299 = .0.72629.*

2. Fit a logistic regression model investigating prevalence of infarcts as a function of age (modeled continuously), coronary heart disease (modeled as dummy variables), and their

multiplicative interaction. Provide a scientific interpretation of each of the regression coefficients, including a description of the intercept in the model. (You do not need to describe the methods, or provide CI or p values.)

***Instructions for grading*: This problem is worth 30 points. I provide answers as if the dummy variables had been fit using chd=0 as the reference group (so reference group is no prior history of angina or MI). I am guessing that no student used any other group as the reference group (and, indeed, I believe it definitely advantageous not to). But if they did, I would expect the grader to ascertain that their answers corresponded to that alternative approach. Similarly, I am guessing that no students would have recoded age. I do expect all students to understand how the re-parameterization of the model affects the interpretations, so be sure you read and understand those additional parts given in problem 1.**

***Stata code and output*:**

```
. logistic infarcts i.chd##c.age
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(5)      =      77.52
                                                  Prob > chi2     =     0.0000
Log likelihood = -2355.4389                       Pseudo R2       =     0.0162
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         chd |
          1  |   1.511575   2.390843     0.26   0.794     .0680916    33.55568
          2  |   10.07338   16.93543     1.37   0.169     .3733623    271.7817
             |
         age |   1.048386   .0080074     6.19   0.000     1.032808    1.064198
             |
   chd#c.age |
          1  |   .9978295    .020616    -0.11   0.916     .9582301    1.039065
          2  |   .9772907   .0216461    -1.04   0.300     .9357728    1.020651
------------------------------------------------------------------------------

. logit infarcts i.chd##c.age
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(5)      =      77.52
                                                  Prob > chi2     =     0.0000
Log likelihood = -2355.4389                       Pseudo R2       =     0.0162
------------------------------------------------------------------------------
    infarcts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         chd |
          1  |   .4131525   1.581689     0.26   0.794    -2.686901    3.513206
          2  |   2.309897   1.681206     1.37   0.169    -.9852061    5.604999
             |
         age |   .0472514   .0076379     6.19   0.000     .0322815    .0622214
             |
   chd#c.age |
          1  |  -.0021728   .0206609    -0.11   0.916    -.0426674    .0383218
          2  |  -.0229712   .0221491    -1.04   0.300    -.0663826    .0204403
             |
       _cons |  -4.350314     .57629    -7.55   0.000    -5.479821   -3.220806
------------------------------------------------------------------------------

. di exp(-4.350314)
.01290276
```

Ans: **Interpretations of the various regression parameters are as follows:**

- **(intercept) In extrapolating from the data on 65 - 100 year olds, the odds of prevalent infarcts in newborns without prior history of angina or MI is estimated to be 0.0129.** *(This is so far outside the range of our data as to be totally irrelevant.)*

- **(slope for age) When comparing groups of two different ages but both groups having no prior history of coronary heart disease, we estimate that the odds of prevalent infarcts is 4.84% higher (OR = 1.0484) for every year that one group is older than the other.**
- **(slope for chd=1) In extrapolating from the data on 65 - 100 year olds, when comparing a group of newborns who have a prior history of angina (but without prior history of MI) to a group of newborns who have no prior history of either angina or MI, we estimate that the odds of prevalent infarcts is 51.2% higher (OR = 1.512) in the group of newborns with prior history of angina.**
- **(slope for chd=2) In extrapolating from the data on 65 - 100 year olds, when comparing a group of newborns who have a prior history of MI (with or without prior history of angina) to a group of newborns who have no prior history of either angina or MI, we estimate that the odds of prevalent infarcts is 10.1-fold higher (OR = 10.1) in the group of newborns with prior history of MI.**
- **(slope for age*chd=1)** *(I provide alternative interpretations)*
  - **When comparing groups of two different ages that are one year apart, the odds ratio for prevalence of infarcts if both groups have prior history of angina (but without prior history of MI) is 0.22% lower than the corresponding OR if both groups have no prior history of either angina or MI.**
  - **When comparing a group of subjects with a prior history of angina (but without prior history of MI) to a group of subjects without prior history of either angina or MI, a comparison made in subjects of a specified age will have an odds ratio for prevalence of infarcts that is 0.22% lower than the corresponding OR for subjects who are 1 year younger.**
- **(slope for age*chd=2)** *(I provide alternative interpretations)*
  - **When comparing groups of two different ages that are one year apart, the odds ratio for prevalence of infarcts if both groups have prior history of MI is 2.27% lower than the corresponding OR if both groups have no prior history of either angina or MI.**
  - **When comparing a group of subjects with a prior history of MI to a group of subjects without prior history of either angina or MI, a comparison made in subjects of a specified age will have an odds ratio for prevalence of infarcts that is 2.27% lower than the corresponding OR for subjects who are 1 year younger.**

3. Fit a logistic regression model that investigates the linearity of the association between the log odds of presence of infarcts and age, after adjustment for coronary heart disease. (Here you do need to describe your methods and results as they relate to the specific question.)

***Instructions for grading*: This problem is worth 10 points. There are an infinite number of ways this could be explored. I present three common ways that might have been considered for defining a more flexible model that includes linearity as a special case. (We should not go overboard in the number of regression parameters are used to model the flexible relationship, or we lose power.) The grader may have to reproduce the methods to be ensure a correct answer. If you cannot reproduce the results, then the description of the Methods may be inadequate. I present results for the three methods of:**

- ***Polynomial regression: I use a cubic polynomial, but a quadratic was not unreasonable.***
- ***Dummy variables: I use four intervals, but other choices are reasonable.***
- ***Linear splines: I use three knots, but other choices are reasonable***

For each of the methods, I present my Stata code and output. A student should only present one approach, and the student should not have presented their Stata output.

Ans: *(Polynomial regression)***:**

Methods: **A logistic regression model of the response variable indicating prevalence of infarct-like lesions was fit to dummy variables modeling prior history of coronary heart disease (no prior history, history of angina only, or history of MI) and age modeled as a cubic polynomial (terms for age, age squared, and age cubed). A test for linearity of the association between log odds of infarct-like lesions and age was performed using a multiple partial chi square test that the regression coefficients for the squared and cubed age terms were simultaneously equal to 0. Statistical significance was defined as a p value less than 0.05.**

```
. g agesqr= age^2
. g agecub= age^3
. logistic infarcts age agesqr agecub i.chd

Logistic regression                                     Number of obs   =       3775
                                                        LR chi2(5)      =      76.74
                                                        Prob > chi2     =     0.0000
Log likelihood = -2355.8336                             Pseudo R2       =     0.0160
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    2.10746   4.836548     0.32   0.745     .0234575    189.3378
      agesqr |   .9915329   .0289901    -0.29   0.771     .9363107    1.050012
      agecub |   1.000034   .0001238     0.28   0.783     .9997916    1.000277
             |
         chd |
          1  |   1.284916   .1443013     2.23   0.026     1.031054    1.601283
          2  |   1.765167   .2009348     4.99   0.000     1.412182    2.206383
------------------------------------------------------------------------------

. test agesqr agecub

 ( 1)  [infarcts]agesqr = 0
 ( 2)  [infarcts]agecub = 0

           chi2(  2) =    0.28
         Prob > chi2 =    0.8693
```

Results: **In a model that adjusted for prior history of CHD, the test that the squared and cubic age term coefficients were both equal to 0 was not statistically significant (P = 0.869), and thus we have insufficient evidence to establish that the association between prevalence of infarct-like lesions and age is nonlinear on the log odds scale.**

*(Dummy variable regression)***:**

Methods: **A logistic regression model of the response variable indicating prevalence of infarct-like lesions was fit to dummy variables modeling prior history of coronary heart disease (no prior history, history of angina only, or history of MI) and age modeled as an untransformed linear variable, along with dummy variables fit to the intervals 65 – 70, 71-76, 77 – 82, and 83 – 100. A test for linearity of the association between log odds of infarct-like lesions and age was performed using a multiple partial chi square test that the regression coefficients for the dummy variable terms were simultaneously equal to 0. Statistical significance was defined as a p value less than 0.05.**

```
. egen agectg= cut(age), at(0,71,77,83,100)
. logistic infarcts age i.agectg i.chd

Logistic regression                                     Number of obs   =       3775
                                                        LR chi2(6)      =      77.62
                                                        Prob > chi2     =     0.0000
```

```
Log likelihood =  -2355.392                     Pseudo R2       =     0.0162
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.05113   .0211863     2.47   0.013     1.010415    1.093486
             |
      agectg |
          71 |    .964295   .1257803    -0.28   0.780     .7467603    1.245198
          77 |   .9972594   .2300162    -0.01   0.991     .6345705    1.567243
          83 |   .8522685   .3083349    -0.44   0.659     .4193991     1.73191
             |
         chd |
           1 |   1.284686   .1442194     2.23   0.026     1.030958     1.60086
           2 |   1.770113    .201516     5.02   0.000      1.41611     2.21261
------------------------------------------------------------------------------

. testparm i.agectg

 ( 1)  [infarcts]71.agectg = 0
 ( 2)  [infarcts]77.agectg = 0
 ( 3)  [infarcts]83.agectg = 0

           chi2(  3) =    1.16
         Prob > chi2 =    0.7615
```

Results: **In a model that adjusted for prior history of CHD, the test that the dummy variable coefficients were all equal to 0 was not statistically significant (P = 0.762), and thus we have insufficient evidence to establish that the association between prevalence of infarct-like lesions and age is nonlinear on the log odds scale.**

*(Linear spline regression)***:**

Methods: **A logistic regression model of the response variable indicating prevalence of infarct-like lesions was fit to dummy variables modeling prior history of coronary heart disease (no prior history, history of angina only, or history of MI) and age modeled as linear splines fit to the intervals 65 – 71, 71-77, 77 – 83, and 83 – 100. A test for linearity of the association between log odds of infarct-like lesions and age was performed using a multiple partial chi square test that the regression coefficients for the linear spline terms were all equal to each other. Statistical significance was defined as a p value less than 0.05.** *(Note that I actually effected the test by including the linear term and then testing that coefficients for the linear spline terms (that were not omitted due to collinearity) were equal to 0. This is exactly the same test that I described, just parameterized in a different way. I could therefore describe it either way.)*

```
. mkspline sage65 71 sage71 77 sage77 83 sage83 = age
. logistic infarcts age sage* i.chd
note: sage83 omitted because of collinearity

Logistic regression                               Number of obs   =       3775
                                                  LR chi2(6)      =      76.80
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2355.803                       Pseudo R2       =     0.0160
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.029418    .034925     0.85   0.393     .9631923    1.100197
       sage65 |    1.03164    .061251     0.52   0.600     .9183123    1.158954
       sage71 |   1.014675   .0378987     0.39   0.696      .943049    1.091742
       sage77 |    1.01954   .0524013     0.38   0.707     .9218394    1.127596
       sage83 |  (omitted)
              |
         chd |
           1 |    1.28534   .1444378     2.23   0.025     1.031254    1.602029
           2 |   1.764323   .2008741     4.99   0.000     1.411451    2.205414
------------------------------------------------------------------------------

. testparm sage*
```

```
 ( 1)  [infarcts]sage65 = 0
 ( 2)  [infarcts]sage71 = 0
 ( 3)  [infarcts]sage77 = 0

           chi2(  3) =    0.34
         Prob > chi2 =    0.9520
```

Results: **In a model that adjusted for prior history of CHD, the test that the linear spline coefficients were all equal was not statistically significant (P = 0.952), and thus we have insufficient evidence to establish that the association between prevalence of infarct-like lesions and age is nonlinear on the log odds scale.**

4. Fit a logistic regression model that investigates whether there is a U-shaped association between the log odds of presence of infarcts and ldl, after adjustment for age. (Here you do need to describe your methods and results as they relate to the specific question.)

Instructions for grading: This problem is worth 10 points. There are an infinite number of ways this could be explored. I present two common ways that might have been considered for defining a more flexible model that allows testing of the (somewhat vague) hypothesis that the association might be U-shaped. (We should not go overboard in the number of regression parameters are used to model the flexible relationship, or we lose power.) The grader may have to reproduce the methods to be ensure a correct answer. If you cannot reproduce the results, then the description of the Methods may be inadequate. I present results for the three methods of:

- ***Linear splines: In this approach I consider whether the slope for the lowest values of LDL is significantly different from zero and of opposite sign of the slope for the highest values of LDL. I use two knots, but other choices are possible. However, I do note that the more intervals, the more variable are the estimates. Hence, choosing which of the slopes to consider is more difficult. With only two knots it is straightforward to specify the test.***
- ***Dummy variables: In this approach I consider whether the coefficients for the lowest and highest intervals are both significantly different in the same direction from the fitted values for a middle interval. Other choices are possible, but as with the linear splines, choosing more intervals leads to greater variability of the estimates and makes specification of the test more difficult.***

For each of the methods, I present my Stata code and output. A student should only present one approach, and the student should not have presented their Stata output.

Ans: *(Linear spline regression)***:**

Methods: **A logistic regression model of the response variable indicating prevalence of infarct-like lesions was fit to age modeled as an untransformed linear predictor and LDL fit as linear splines fit to the intervals 13 - 100, 100 – 160, and 160 - 413. A test for a generally U-shaped association between log odds of infarct-like lesions and LDL was performed by testing that the slopes for the lowest and highest LDL intervals were both significantly different from 0 and that the two slopes were opposite in sign. Statistical significance was defined as a p value less than 0.05.** *(Note that this approach essentially defines what we mean by a U-shaped function. You could imagine that if I had fit more intervals, some sort of a sinusoidal pattern might also meet these criteria. That is of course a far more complicated pattern to describe and to have precision to test. I strongly urge building up your knowledge about the "dose-response" slowly. I could have just fit two intervals and looked for the signs of the slopes to be significantly different from zero and opposite in sign. However, the possibility that I might have guessed the wrong threshold for the change in sign makes three intervals seem slightly better to me. Using only three*

intervals seems a reasonable tradeoff between having precision to estimate the trend when any true "nadir" or "maximum" is unknown and having sufficient to estimate the slopes.)

```
. mkspline sldl0 100 sldl00 160 sldl160= ldl
. logistic infarcts sldl* age

Logistic regression                               Number of obs   =       3731
                                                  LR chi2(4)      =      53.86
                                                  Prob > chi2     =     0.0000
Log likelihood = -2339.7353                       Pseudo R2       =     0.0114
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sldl0 |   .9994646   .0035642    -0.15   0.881     .9925033    1.006475
      sldl00 |   1.000525   .0019973     0.26   0.793     .9966181    1.004447
     sldl160 |   1.009152   .0035372     2.60   0.009     1.002243    1.016109
         age |   1.046296   .0070583     6.71   0.000     1.032553    1.060222
------------------------------------------------------------------------------
```

Results: **In a model that adjusted for age, the logistic regression slope describing the prevalence of infarct-like lesions in the highest interval of LDL was significantly different from 0 and in a positive direction (OR = 1.0091 when comparing individuals differing by 1 mg/dL over the interval 160 – 413 mg/dL, P = 0.009), however the corresponding regression slope in the lowest interval of LDL, though estimated to be opposite in sign to that of the highest LDL interval, was not statistically significantly different from 0 (OR = 0.9995 when comparing individuals differing by 1 mg/dL over the interval 13 - 100 mg/dL, P = 0.881). Thus we have insufficient evidence to establish that the association between prevalence of infarct-like lesions and LDL is U-shaped on the log odds scale.**

*(Dummy variable regression)***:**

Methods: **A logistic regression model of the response variable indicating prevalence of infarct-like lesions was fit to age modeled as an untransformed linear predictor and LDL fit as dummy variables to the intervals 13 - 99, 100 – 159, and 160 - 413. A test for a generally U-shaped association between log odds of infarct-like lesions and LDL was performed by testing that the fitted values for the lowest and highest LDL intervals were both significantly different from that for the intermediate levels of LDL and that the two extreme intervals fitted values were both higher or both lower than those for the intermediate intervals. Statistical significance was defined as a p value less than 0.05.** *(Note that this approach essentially defines what we mean by a U-shaped function. I had to fit at least three intervals for this approach. You could imagine that if I had fit more intervals, some sort of a sinusoidal pattern might also meet these criteria. That is of course a far more complicated pattern to describe and to have precision to test. I strongly urge building up your knowledge about the "dose-response" slowly. Using only three intervals seems a reasonable tradeoff between having precision to estimate the trend when any true "nadir" or "maximum" is unknown and having sufficient to estimate the slopes.)*

```
. egen ldlctg= cut(ldl), at(0,100,160,450)
(44 missing values generated)
. logistic infarcts ib100.ldlctg age

Logistic regression                               Number of obs   =       3731
                                                  LR chi2(3)      =      51.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -2341.0972                       Pseudo R2       =     0.0108
------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ldlctg |
           0 |   1.071625   .0889738     0.83   0.405     .9106893    1.261001
         160 |   1.318956   .1392302     2.62   0.009     1.072449    1.622123
------------------------------------------------------------------------------
```

```
             |
         age |      1.046   .0070462     6.68   0.000      1.03228    1.059902
```

Results: **In a model that adjusted for age, the logistic regression slope describing the prevalence of infarct-like lesions in the highest interval of LDL (160 – 413 mg/dL) was significantly different from that in the intermediate range (100 – 159 mg/dL) and in a positive direction (OR = 1.319, P = 0.009), however the corresponding fitted values in the lowest interval of LDL (13 – 99 mg/dL), though estimated to be also higher than the intermediate interval, the difference was not statistically significantly different from 0 (OR =1.072, P = 0.405). Thus we have insufficient evidence to establish that the association between prevalence of infarct-like lesions and LDL is U-shaped on the log odds scale.**