# Biost 536 / Epi 536
# Categorical Data Analysis in Epidemiology

## Midterm Examination Key
## November 7, 2013

Name: _____

**Instructions: This exam is closed book, closed notes. You have 90 minutes. You may not use any device that is capable of accessing the internet.**

**Please provide concise answers to all questions. Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.**

**NOTE: When you need to make calculations, always use at least four significant digits in your intermediate calculations, and report at least three significant digits. (Example: 1.045 and 0.0001234 and 1234000 each have four significant digits.) Give some indication of how you were calculating your answer. (If you give the wrong answer, but I can determine where you went wrong, you may get partial credit.)**

**If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.**

**Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor.**

**PLEDGE:**
**On my honor, I have neither given nor received unauthorized aid on this examination:**

**Signed:**_____

All problems deal with a subset of data from an observational study of serum cholesterol and mortality in an elderly population. The appendices contain results from selected analyses:

Appendix A : Description of the variables and descriptive statistics **(all problems)**
Appendix B : Analyses of 5 year mortality by dichotomized LDL **(problems 1 through 6)**
Appendix C : Logistic regression analysis of 5 year mortality by categorized LDL **(problem 7)**
Appendix D : Logistic regression analysis of 5 year mortality by quadratic LDL, sex, age

1. Appendix B provides data on the 5 year mortality within groups defined by LDL less than or greater than 160 mg/dL and sex. Suppose we let $p$ be the 5 year mortality, and we are interested in performing **linear regression** involving main effects for ***ldlGE160*** and ***male*** and a multiplicative interaction ***ldlGE160_male = ldlGE160*** x ***male*** .

$$p = \beta_0 + \beta_1 \times ldlGE160 + \beta_2 \times male + \beta_3 \times ldlGE160\_male$$

*(The solution to this problem uses the facts that*

- *we are modeling the risk of death within 4 years, and*
- *this is a saturated model that models the risk in four groups using four parameters,*
- *hence the fitted values will correspond exactly to the sample proportions.)*

a. Can you calculate the estimated ***intercept*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: 0.1221**

*(The intercept is the risk of death within 4 years among females having low LDL.)*

b. Can you calculate the estimated slope for ***ldlGE160*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: 0.0806 – 0.1221 = -0.0415**

*(This slope is the difference in the risk of death within 4 years for females having high LDL and the risk of death within 4 years for females having low LDL.)*

c. Can you calculate the estimated slope for ***male*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: 0.2159 – 0.1221 = 0.0938**

*(This slope is the difference in the risk of death within 4 years for males having low LDL and the risk of death within 4 years for females having low LDL.)*

d. Can you calculate the estimated slope for the interaction ***ldlGE160_male*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: (0.2000 – 0.2159) – (0.0806 – 0.1221) = (0.2000 – 0.0806) – (0.2159 – 0.1221) = 0.0256.**

*(This slope is the difference in the risk difference across LDL levels for males and the risk difference across LDL levels for females.*

*Alternatively, this slope is the difference in the risk difference across sexes for low LDL subjects and the risk difference across sexes levels for low LDL levels.)*

2. Appendix B provides data on the 5 year mortality within groups defined by LDL less than or greater than 160 mg/dL and sex. Suppose we let $p$ be the 5 year mortality, and we are interested in performing **Poisson regression** involving main effects for ***ldlGE160*** and ***male*** and a multiplicative interaction ***ldlGE160_male = ldlGE160*** x ***male*** .

$$\log(p) = \beta_0 + \beta_1 \times ldlGE160 + \beta_2 \times male + \beta_3 \times ldlGE160\_male$$

*(The solution to this problem uses the facts that*

- *we are modeling the log risk of death within 4 years, and*
- *this is a saturated model that models the risk in four groups using four parameters,*
- *hence the fitted values will correspond exactly to the logarithm of sample proportions.)*

a. Can you calculate the estimated ***intercept*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log (0.1221) = -2.103**

*(The intercept is the log risk of death within 4 years among females having low LDL.)*

b. Can you calculate the estimated slope for ***ldlGE160*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log (0.0806 / 0.1221) = log (0.0806) – log (0.1221) = -0.4153**

*(This slope is the log of the risk ratio (RR) comparing risk of death within 4 years for females having high LDL and the risk of death within 4 years for females having low LDL.*

*Equivalently, this slope is the difference of the logarithm of the risk of death within 4 years for females having high LDL and the logarithm of the risk of death within 4 years for females having low LDL )*

c. Can you calculate the estimated slope for ***male*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log (0.2159 / 0.1221) = log (0.2159) – log (0.1221) = 0.5700**

*(This slope is the log of the risk ratio (RR) comparing risk of death within 4 years for males having low LDL and the risk of death within 4 years for females having low LDL.*

*Equivalently, this slope is the difference of the logarithm of the risk of death within 4 years for males having low LDL and the logarithm of the risk of death within 4 years for females having low LDL )*

d. Can you calculate the estimated slope for the interaction ***ldlGE160_male*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log ( (0.2000 / 0.2159) / (0.0806 / 0.1221) ) = 0.3388**

*(This slope is the logarithm of the ratio of the risk ratio across LDL levels for males and the risk ratio across LDL levels for females. Equivalently, it is the difference of the logarithm of the risk ratio across LDL levels for males and the logarithm of the risk ratio across LDL levels for females. And we could also describe it as the differences of differences of the log risk. You get the picture.*

*And we again have an alternative interpretation of the interaction slope as the logarithm of the ratio of the risk ratio across sexes for subjects with high LDL levels and the risk ratio across sexes for subjects with low LDL levels.)*

3. Appendix B provides data on the 5 year mortality within groups defined by LDL less than or greater than 160 mg/dL and sex. Suppose we let $p$ be the 5 year mortality, and we are interested in performing **logistic regression** involving main effects for ***ldlGE160*** and ***male*** and a multiplicative interaction ***ldlGE160_male = ldlGE160*** x ***male*** .

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times ldlGE160 + \beta_2 \times male + \beta_3 \times ldlGE160\_male$$

*(The solution to this problem uses the facts that*

- *we are modeling the log odds of death within 4 years, and*
- *this is a saturated model that models the risk in four groups using four parameters,*
- *hence the fitted values will correspond exactly to the logarithm of the sample odds.)*

a. Can you calculate the estimated ***intercept*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log (0.1221 / (1 – 0.1221) ) = -1.973**

*(The intercept is the log risk of death within 4 years among females having low LDL.)*

b. Can you calculate the estimated slope for ***ldlGE160*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log ( (0.0806 / (1 – 0.0806) ) / ( 0.1221 / (1 – 0.1221) ) ) = -0.4615**

*(This slope is the log of the odds ratio (OR) comparing odds of death within 4 years for females having high LDL and the odds of death within 4 years for females having low LDL. We could also have described it as the difference of log odds.)*

c. Can you calculate the estimated slope for ***male*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log ( (0.2159 / (1 – 0.2159) ) / (0.1221 / (1 – 0.1221) ) ) = 0.6830**

*(This slope is the log of the odds ratio (OR) comparing odds of death within 4 years for males having low LDL and the odds of death within 4 years for females having low LDL. We could also have described it as the difference of log odds.)*

d. Can you calculate the estimated slope for the interaction ***ldlGE160_male*** that would be obtained from that regression? If so provide the estimate. If not, explain why not.

**Ans: log(((0.2000/(1–0.2000))/(0.2159/(1–0.2159)))/((0.0806/(1–0.0806) )/( 0.1221/(1–0.1221))) = 0.3650**

*(This slope is the logarithm of the ratio of the odds ratio across LDL levels for males and the odds ratio across LDL levels for females. Equivalently, it is the difference of the logarithm of the odds ratio across LDL levels for males and the logarithm of the odds ratio across LDL levels for females.*

*And we again have an alternative interpretation of the interaction slope as the logarithm of the ratio of the odds ratio across sexes for subjects with high LDL levels and the odds ratio across sexes for subjects with low LDL levels.)*

4. Appendix B provides data on the 5 year mortality within groups defined by LDL less than or greater than 160 mg/dL and sex. Suppose we let $p$ be the 5 year mortality, and we are interested in performing **logistic regression** involving main effects for ***ldlGE160*** and ***male*** (without the interaction) . Can you similarly calculate what the regression coefficients would be for this model? If so provide the estimates. If not, explain why not.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times ldlGE160 + \beta_2 \times male$$

**Ans: I cannot do this, because we do not have a saturated model and the parameter estimates will end up borrowing data across strata. For instance, the slope for male will be some sort of weighted average of the log odds ratio comparing males with low LDL to females with low LDL and the log odds ratio comparing males with high LDL to females with high LDL. The exact weighting is found in an iterative computer search.**

*(Had I asked this question about linear regression, it would be possible to figure out the weighting based on harmonic means of sample sizes. I discussed this in the annotated Stata output for homework #2. I certainly would not have expected you to remember that formula,)*

5. Again using Appendices B, does sex confound the estimation of an association between LDL and 5 year mortality?

a. Answer the question assuming you are using risk difference as a measure of association.

**Ans: We can first consider whether there is an association between LDL level and sex in our sample. We can look at this either of two ways:**

- **Among females, 62 / 365 = 17.0% have high LDL measurements, while among males, 45 / 360 = 12.5% have high LDL measurements.**

- **Among subjects with high LDL, 45 / 107 = 42.05% are male, while among subjects with low LDL, 315 / 618 = 50.97% are male.**

**If we regard that observed differences in the prevalence of high LDL or the observed differences in the sex ratio as unimportant, then we can immediately say there is no confounding.** *(Note that each of the above comparisons are just different ways of describing the same association. Had I used OR as a measure of association, there would not have been an issue in deciding which way to look at it.)*

**If we regard that the difference in prevalence is important, the we have to consider whether sex is causally associated with 5 year mortality within groups defined by LDL level. *A priori* we know that males do not survive as well as females for reasons beyond lipid profiles. This is borne out in the data: Among subjects with low LDL, the 5 year mortality for males is 21.59% and for females is 12.21%. Similarly, Among subjects with high LDL, the 5 year mortality for males is 20.00% and for females is 8.06%. I would regard these differences as evidence of an association.**

*(So long as you correctly justified your answer, you would receive full credit no matter whether you said yes or no.)*

b. Answer the question assuming you are using risk ratio as a measure of association.

**Ans: In this setting of two binary predictors, our answer would not vary whether we considered RD, RR, or OR.**

c. Answer the question assuming you are using odds ratio as a measure of association.

**Ans: In this setting of two binary predictors, our answer would not vary whether we considered RD, RR, or OR.**

6. Again using Appendix B, does sex modify any association between LDL and 5 year mortality?

a. Answer the question assuming you are using risk difference as a measure of association.

**Ans: The 5 year mortality for males with low LDL is 21.59%, and the 5 year mortality for males with high LDL is 20.00%. The risk difference is RD = -1.59%.**

**The 5 year mortality for females with low LDL is 12.21%, and the 5 year mortality for females with high LDL is 8.06%. The risk difference is RD = -4.15%.**

**If we regard the difference between -1.59% and -4.15% important, then there is effect modification. If we do not, then there is no effect modification.** *(I note that in trying to*

*judge whether differences of 2.56% in 5 year mortality is important, we could consider that the RD comparing sexes is 9.66%, so the magnitude of the observed effect modification is about one-fourth the magnitude of an (unadjusted) association between mortality and sex. Though not shown here, that magnitude is comparable to the RD corresponding to about a 2.5 year difference in age. You merely had to be justifying your answer correctly.)*

b. Answer the question assuming you are using risk ratio as a measure of association.

**Ans: The 5 year mortality for males with low LDL is 21.59%, and the 5 year mortality for males with high LDL is 20.00%. The risk ratio is RR = 0.926.**

**The 5 year mortality for females with low LDL is 12.21%, and the 5 year mortality for females with high LDL is 8.06%. The risk ratio is RR = 0.660.**

**I regard the difference between 0.926 and 0.660 sufficient to regard that there is effect modification on the risk ratio scale.** *(You merely had to be justifying your answer correctly For what it is worth, we do not have sufficient precision to regard.)*

c. Answer the question assuming you are using odds ratio as a measure of association.

**Ans: The 5 year mortality for males with low LDL is 21.59% (odds = 0.2753), and the 5 year mortality for males with high LDL is 20.00% (odds = 0.2500). The odds ratio is OR = .9081**

**The 5 year mortality for females with low LDL is 12.21% (odds = 0.1391), and the 5 year mortality for females with high LDL is 8.06% (odds = 0.0877). The odds ratio is OR = 0.6305.**

**I regard the difference between 0.908 and 0.631 sufficient to regard that there is effect modification on the odds ratio scale.** *(You merely had to be justifying your answer correctly.)*

7. Appendix C contains the results of unadjusted and age adjusted logistic regression analyses where LDL is modeled using using categorical variables

a. In the **unadjusted** model what is the interpretation of the intercept?

**Ans: The log odds of mortality within 5 years is estimated to be -.3677 among subjects having LDL less than 70 mg/dL.**

**Hence, the odds of mortality within 5 years is estimated to be $e^{-.3677}$ = 0.6923 among subjects having LDL less than 70 mg/dL.**

**Hence, the probability of mortality within 5 years is estimated to be 0.6923 / 1.6923 = 0.4091 among subjects having LDL less than 70 mg/dL.**

*(Any one of the three interpretations is adequate. My personal preference would be the odds if I am also going to interpret the slopes as OR.)*

b. In the **unadjusted** model what is the interpretation of the slope for level 2?

**Ans: The log odds ratio comparing the odds of death within 5 years for subjects having LDL between 70 and 100 mg/dL to the odds of death within 5 years for subjects having LDL less than 70 mg/dL is estimated to be -1.233.**

**Hence, the odds ratio comparing the odds of death within 5 years for subjects having LDL between 70 and 100 mg/dL to the odds of death within 5 years for subjects having LDL less than 70 mg/dL is estimated to be $e^{-1.233} = 0.2913$.**

*(Again, either one of the above is acceptable, but my preference would be to give the OR.)*

c. In the **unadjusted** model, can you provide a p value for an association between LDL and 5 year mortality? If so, do so. If not, explain what you would want instead.

**Ans: Because every modeled covariate pertains to LDL, using this model we would judge that there is no association between mortality and LDL if and only if every covariate in the model had a zero coefficient. This is equivalent to comparing the fitted model to a logistic regression model that had no covariates. The likelihood ratio statistic makes such a comparison, and that P value is 0.0374. (Hence, using a two-sided level 0.05 level of significance, we would reject the null hypothesis of no association.)**

*(The likelihood ratio statistic is equivalent to the Wald test in large samples, but in small samples, there will typically be differences between the P values obtained. In this analysis, the Wald test provides a P value of 0.0282 and is obtained in Stata as follows:*

```
. testparm i.ldlctg

 ( 1)  [deadin5]2.ldlctg = 0
 ( 2)  [deadin5]3.ldlctg = 0
 ( 3)  [deadin5]4.ldlctg = 0
 ( 4)  [deadin5]5.ldlctg = 0
 ( 5)  [deadin5]6.ldlctg = 0

           chi2(  5) =   12.53
         Prob > chi2 =    0.0282

. test i2.ldlctg i3.ldlctg i4.ldlctg i5.ldlctg i6.ldlctg

 ( 1)  [deadin5]2.ldlctg = 0
 ( 2)  [deadin5]3.ldlctg = 0
 ( 3)  [deadin5]4.ldlctg = 0
 ( 4)  [deadin5]5.ldlctg = 0
 ( 5)  [deadin5]6.ldlctg = 0

           chi2(  5) =   12.53
         Prob > chi2 =    0.0282
```

*As a general rule, we trust the likelihood ratio test more than the Wald test in small samples.)*

d. In the **age adjusted model** model what is the interpretation of the intercept?

**Ans: The log odds of mortality within 5 years is estimated to be -5.974 among newborns having LDL less than 70 mg/dL.**

**Hence, the odds of mortality within 5 years is estimated to be $e^{-5.974}$ = 0.002544 among newborns having LDL less than 70 mg/dL.**

**Hence, the probability of mortality within 5 years is estimated to be 0.002544 / 1.002544 = 0.002538 among newborns having LDL less than 70 mg/dL.**

*(Any one of the three interpretations is adequate. My personal preference would be the odds if I am also going to interpret the slopes as OR. This is extrapolating way outside the range of our data, so I would make no scientific use of this data.)*

e. In the **age adjusted model** model what is the interpretation of the slope for level 2?

**Ans: The log odds ratio comparing the odds of death within 5 years for subjects having LDL between 70 and 100 mg/dL to the odds of death within 5 years for subjects of the same age having LDL less than 70 mg/dL is estimated to be -1.214.**

**Hence, the odds ratio comparing the odds of death within 5 years for subjects having LDL between 70 and 100 mg/dL to the odds of death within 5 years for subjects of the same age having LDL less than 70 mg/dL is estimated to be $e^{-1.214}$ = 0.2970.**

*(Again, either one of the above is acceptable, but my preference would be to give the OR.)*

f. In the **age adjusted model**, can you provide a p value for an association between LDL and 5 year mortality? If so, do so. If not, explain what you would want instead.

**Ans: In this regression model, there is one covariate that models the LDL-adjusted association between age and 5 year mortality, and there are five covariates that in some way model the age-adjusted association between LDL and 5 year mortality. In order to test the age-adjusted association between LDL and 5 year mortality, we would need each of the regression parameters for the variables i2.ldlctg, i3.ldlctg, i4.ldlctg, i5.ldlctg, i6.ldlctg to be 0. So we would need to simultaneously test all of those parameters.**

*In Stata, we could use the* `test` *or* `testparm` *commands to perform Wald tests:*

```
. test i2.ldlctg i3.ldlctg i4.ldlctg i5.ldlctg i6.ldlctg

 ( 1)  [deadin5]2.ldlctg = 0
 ( 2)  [deadin5]3.ldlctg = 0
 ( 3)  [deadin5]4.ldlctg = 0
 ( 4)  [deadin5]5.ldlctg = 0
 ( 5)  [deadin5]6.ldlctg = 0

           chi2(  5) =   11.74
         Prob > chi2 =    0.0386
```

```
. testparm i.ldlctg

 ( 1)  [deadin5]2.ldlctg = 0
 ( 2)  [deadin5]3.ldlctg = 0
 ( 3)  [deadin5]4.ldlctg = 0
 ( 4)  [deadin5]5.ldlctg = 0
 ( 5)  [deadin5]6.ldlctg = 0

           chi2(  5) =   11.74
         Prob > chi2 =    0.0386
```

*Alternatively, we could first store the results from the current model, then fit a "reduced" model that has only the age covariate (making sure that we use only the same cases that were used in the "full model"—several cases were missing LDL), and then perform a likelihood ratio test:*

```
. est store fullModel

. logit deadin5 age

Logistic regression                               Number of obs   =        735
                                                  LR chi2(1)      =      16.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -320.37454                       Pseudo R2       =     0.0254
------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |     .07065   .0170736     4.14   0.000     .0371864    .1041136
       _cons |  -6.941961   1.300616    -5.34   0.000    -9.491122   -4.392801
------------------------------------------------------------------------------

. lrtest fullModel
observations differ: 725 vs. 735
r(498);

. logit deadin5 age if ldl!=.

Logistic regression                               Number of obs   =        725
                                                  LR chi2(1)      =      18.15
                                                  Prob > chi2     =     0.0000
Log likelihood =  -314.6124                       Pseudo R2       =     0.0280
------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0741496   .0172167     4.31   0.000     .0404054    .1078937
       _cons |  -7.211568   1.312632    -5.49   0.000    -9.784279   -4.638857
------------------------------------------------------------------------------

. lrtest fullModel

Likelihood-ratio test                                 LR chi2(5)  =     11.13
(Assumption: . nested in fullModel)                   Prob > chi2 =    0.0488
```

## **APPENDIX A**: **Description of variables and descriptive statistics**

These data come from an observational study of blood lipids as measured by serum LDL (the "bad cholesterol") and 5 year mortality in a population of elderly U.S. residents. All subjects were followed for at least 5 years from the time of study enrolment.

Recommendations for risk of cardiovascular disease according to serum LDL (low density lipoprotein) levels are as follows (taken from the Mayo Clinic website):

| | |
|---|---|
| Below 70 mg/dL | Ideal for people at very high risk of heart disease |
| Below 100 mg/dL | Ideal for people at risk of heart disease |
| 100-129 mg/dL | Near ideal |
| 130-159 mg/dL | Borderline high |
| 160-189 mg/dL | High |
| 190 mg/dL and above | Very high |

This exam considers the following variables (all measured at time of study enrolment) on a subset of 732 subjects from that study.

***age***: age in years of the subject at the time of study enrolment
***male***: indicator that the subject is male (**0**= female, **1**= male)
***ldl***: serum LDL for the subject at time of study enrolment (mg/dL).
***deadin5***: indicator that the subject died within 5 years of study enrolment (**0**= alive at 5 years, **1**= died within 5 years)

The following variables were derived from the serum LDL measurements in order to use in some data analyses:

***ldlsqr***: the square of the serum LDL value as generated from Stata command

```
g ldlsqr = ldl^2
```

***ldlGE160***: indicator that the serum LDL measurement is greater than or equal to 160 mg/DL (**0**= *ldl* < 160 mg/dL, **1**= *ldl* ≥ 160 mg/dL) as generated from Stata command:

```
recode ldl 160/max=1 min/160=0, gen(ldlGE160)
```

***ldlctg***: categorized levels of serum LDL using the categories described by the Mayo Clinic:
(**1** = *ldl* < 70 mg/dL; **2**= 70 mg/dL ≤ *ldl* < 100 mg/dL; ; **3**= 100 mg/dL ≤ *ldl* < 130 mg/dL; **4**= 130 mg/dL ≤ *ldl* < 160 mg/dL; **5**= 160 mg/dL ≤ *ldl* < 190 mg/dL; ; **6**= 190 mg/dL ≤ *ldl*) as generated from Stata commands:

```
recode ldl 190/max=16 160/190=5 130/160=4 100/130=3 70/100=2 min/70=1, gen(ldlctg)
```

Variables ***ldl56***, ***ldl87***, ***ldl115***, ***ldl143***, ***ldl172***, and ***ldl208*** modeling linear splines were also created using those same intervals of LDL measurements by the following Stata code

```
mkspline ldl56 70 ldl87 100 ldl115 130 ldl143 160 ldl172 190 ldl208 = ldl
```

(I named the variables according to the average LDL measurement over each range.)

A multiplicative variable ***ldl_male*** was created to model interactions between sex and ldl:

```
g ldl_male= ldl * male
```

**APPENDIX A (cont.)**: **Description of variables and descriptive statistics**

The following table presents descriptive statistics for the above variables within strata defined by LDL categories, as well as for the entire sample. There is no missing data for any variable. Descriptive statistics include the sample size (N), sample mean, sample standard deviation (sd), sample minimum (min), sample 25$^{th}$ percentile (p25), sample median (p50), sample 75$^{th}$ percentile (p75) and sample maximum (max).:

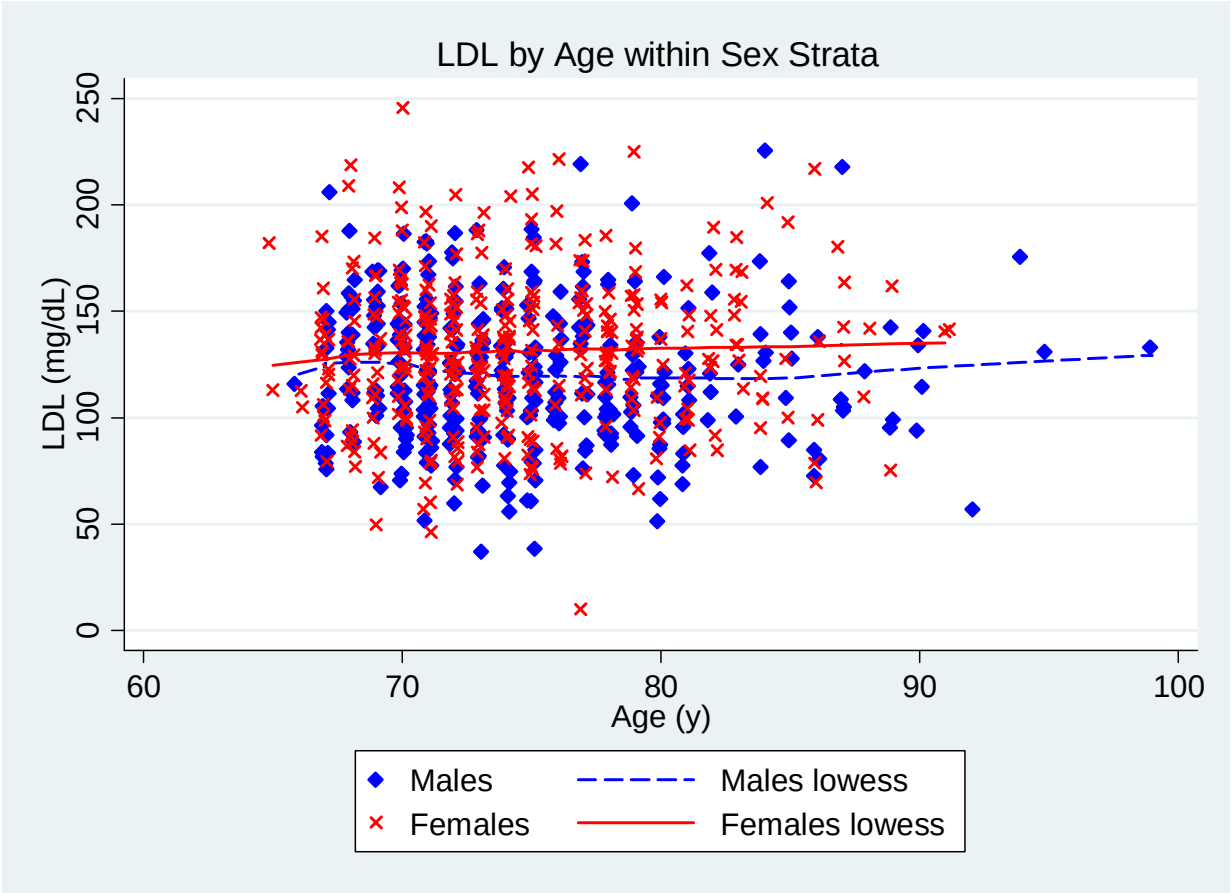
**. tabstat male age ldl deadin5, by(ldlctg) col(stat) stat(n mean sd min q max) long**

| ldlctg | variable | N | mean | sd | min | p25 | p50 | p75 | max |
|---|---|---|---|---|---|---|---|---|---|
| 1 | male | 22 | .6818182 | .4767313 | 0 | 0 | 1 | 1 | 1 |
| | age | 22 | 75.54545 | 5.629237 | 69 | 71 | 74 | 79 | 92 |
| | ldl | 22 | 56.18182 | 13.64389 | 11 | 52 | 60.5 | 67 | 69 |
| | deadin5 | 22 | .4090909 | .5032363 | 0 | 0 | 0 | 1 | 1 |
| 2 | male | 143 | .5594406 | .4981993 | 0 | 0 | 1 | 1 | 1 |
| | age | 143 | 74.65734 | 5.487636 | 67 | 71 | 73 | 78 | 90 |
| | ldl | 143 | 86.68531 | 8.245716 | 70 | 80 | 88 | 94 | 99 |
| | deadin5 | 143 | .1678322 | .3750308 | 0 | 0 | 0 | 0 | 1 |
| 3 | male | 228 | .5394737 | .4995361 | 0 | 0 | 1 | 1 | 1 |
| | age | 228 | 74.64035 | 5.077596 | 65 | 71 | 74 | 78 | 90 |
| | ldl | 228 | 114.7412 | 8.36495 | 100 | 108 | 114 | 122 | 129 |
| | deadin5 | 228 | .1885965 | .3920484 | 0 | 0 | 0 | 0 | 1 |
| 4 | male | 225 | .4311111 | .4963358 | 0 | 0 | 0 | 1 | 1 |
| | age | 225 | 74.19556 | 5.624165 | 67 | 70 | 73 | 77 | 99 |
| | ldl | 225 | 142.7333 | 8.528314 | 130 | 136 | 142 | 150 | 159 |
| | deadin5 | 225 | .1288889 | .335824 | 0 | 0 | 0 | 0 | 1 |
| 5 | male | 83 | .4819277 | .5027108 | 0 | 0 | 0 | 1 | 1 |
| | age | 83 | 74.56627 | 5.668038 | 65 | 70 | 73 | 78 | 94 |
| | ldl | 83 | 172.2771 | 9.214667 | 160 | 164 | 171 | 181 | 189 |
| | deadin5 | 83 | .1204819 | .3275031 | 0 | 0 | 0 | 0 | 1 |
| 6 | male | 24 | .2083333 | .4148511 | 0 | 0 | 0 | 0 | 1 |
| | age | 24 | 75.95833 | 6.111139 | 67 | 70.5 | 75 | 80.5 | 87 |
| | ldl | 24 | 208.3333 | 13.47676 | 191 | 196.5 | 206 | 216.5 | 247 |
| | deadin5 | 24 | .1666667 | .3806935 | 0 | 0 | 0 | 0 | 1 |
| Total | male | 725 | .4965517 | .5003333 | 0 | 0 | 0 | 1 | 1 |
| | age | 725 | 74.56828 | 5.446103 | 65 | 71 | 74 | 78 | 99 |
| | ldl | 725 | 125.8028 | 33.60197 | 11 | 102 | 125 | 147 | 247 |
| | deadin5 | 725 | .1641379 | .3706564 | 0 | 0 | 0 | 0 | 1 |

**APPENDIX A (cont.)**: **Description of variables and descriptive statistics**

The following is a scatterplot of LDL by age, with points plotted according to sex. Superimposed on the plot are lowess curves, again stratified by sex. (Males: blue diamonds and dashed lowess line, Females: red X and solid lowess line)

```
twoway (scatter ldl age if male==1, jitter(1) ms(d) col(blue)) ///
       (lowess ldl age if male==1, col(blue) lp(dash))         ///
       (scatter ldl age if male==0, jitter(1) ms(X) col(red))  ///
       (lowess ldl age if male==0, col(red) lp(solid)),        ///
       t1(LDL by Age within Sex Strata)                        ///
       xtitle("Age (y)") ytitle("LDL (mg/dL)")                 ///
       legend(label(1 Males) label(3 Females)                  ///
              label(2 Males lowess) label(4 Females lowess))
```



## APPENDIX B: 5 year mortality by dichotomized LDL level (LDL < 160 vs LDL ≥ 160) overall and within sex groups

```
. tabulate ldlGE160 deadin5, row chi
```

```
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

 RECODE of |        deadin5
       ldl |         0          1 |     Total
-----------+----------------------+----------
         0 |       513        105 |       618
           |     83.01      16.99 |    100.00
-----------+----------------------+----------
         1 |        93         14 |       107
           |     86.92      13.08 |    100.00
-----------+----------------------+----------
     Total |       606        119 |       725
           |     83.59      16.41 |    100.00

          Pearson chi2(1) =   1.0144   Pr = 0.314
```

```
. bysort male: tabulate ldlGE160 deadin5, row chi
```

```
-> male = 0

 RECODE of |        deadin5
       ldl |         0          1 |     Total
-----------+----------------------+----------
         0 |       266         37 |       303
           |     87.79      12.21 |    100.00
-----------+----------------------+----------
         1 |        57          5 |        62
           |     91.94       8.06 |    100.00
-----------+----------------------+----------
     Total |       323         42 |       365
           |     88.49      11.51 |    100.00

          Pearson chi2(1) =   0.8691   Pr = 0.351
```

```
-> male = 1

 RECODE of |        deadin5
       ldl |         0          1 |     Total
-----------+----------------------+----------
         0 |       247         68 |       315
           |     78.41      21.59 |    100.00
-----------+----------------------+----------
         1 |        36          9 |        45
           |     80.00      20.00 |    100.00
-----------+----------------------+----------
     Total |       283         77 |       360
           |     78.61      21.39 |    100.00

          Pearson chi2(1) =   0.0590   Pr = 0.808
```

## APPENDIX C: Logistic regression analyses of 5 year mortality by categorized LDL and age.

```
. logit deadin5 i.ldlctg

Logistic regression                                     Number of obs   =        725
                                                        LR chi2(5)      =      11.81
                                                        Prob > chi2     =     0.0374
Log likelihood = -317.78312                             Pseudo R2       =     0.0182

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ldlctg |
          2  |  -1.233345   .4879592    -2.53   0.011    -2.189728   -.2769625
          3  |  -1.091431   .4655055    -2.34   0.019    -2.003805    -.179057
          4  |  -1.543094   .4770943    -3.23   0.001    -2.478182   -.6080065
          5  |   -1.62015   .5493021    -2.95   0.003    -2.696762   -.5435372
          6  |  -1.241713   .6985944    -1.78   0.075    -2.610933    .1275067
             |
       _cons |  -.3677247   .4336291    -0.85   0.396    -1.217622    .4821727
------------------------------------------------------------------------------

. logit deadin5 i.ldlctg age

Logistic regression                                     Number of obs   =        725
                                                        LR chi2(6)      =      29.29
                                                        Prob > chi2     =     0.0001
Log likelihood = -309.04527                             Pseudo R2       =     0.0452

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ldlctg |
          2  |    -1.2144   .4972311    -2.44   0.015    -2.188955   -.2398447
          3  |  -1.060054   .4744283    -2.23   0.025    -1.989916   -.1301912
          4  |  -1.501805   .4862915    -3.09   0.002    -2.454918   -.5486907
          5  |  -1.608515   .5589789    -2.88   0.004    -2.704094    -.512937
          6  |  -1.329969   .7112592    -1.87   0.062    -2.724011    .0640734
             |
         age |   .0741065   .0175253     4.23   0.000     .0397576    .1084554
       _cons |  -5.973839   1.397759    -4.27   0.000    -8.713396   -3.234283
------------------------------------------------------------------------------
```

## **APPENDIX D**: Logistic regression of quadratic association of 5 year mortality with LDL.

```
. logistic deadin5 ldl ldlsqr male age

Logistic regression                               Number of obs   =        725
                                                  LR chi2(4)      =      37.44
                                                  Prob > chi2     =     0.0000
Log likelihood = -304.97073                       Pseudo R2       =     0.0578

------------------------------------------------------------------------------
     deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   .9686419   .0143145    -2.16   0.031     .9409883    .9971082
      ldlsqr |   1.000102   .0000571     1.78   0.075     .9999898    1.000214
        male |   1.965183   .4217427     3.15   0.002     1.290414    2.992796
         age |   1.075403   .0188819     4.14   0.000     1.039025    1.113055
------------------------------------------------------------------------------


. test ldl ldlsqr

 ( 1)  [deadin5]ldl = 0
 ( 2)  [deadin5]ldlsqr = 0

           chi2(  2) =    7.35
         Prob > chi2 =    0.0254
```