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

## Second Midterm Examination
## November 13, 2014

Name: _____

**Instructions: This exam is closed book, closed notes. You have 110 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:**_____

**Problems 1-3** consider analyses of the ***prevalence of infarct-like lesions on MRI as a function of smoking behaviors***. These analyses use the dataset assigned for the project, and thus these analyses are similar to some of those you might be considering for the project.

Appendix A : Description of the variables and descriptive statistics **(problems 1 - 4)**
Appendix B : Unadjusted analyses of infarct-like lesions by age **(problem 1)**
Appendix C : Unadjusted analyses of infarct-like lesions by smoking history **(problem 2)**
Appendix C : Analyses of infarct-like lesions by smoking history adjusted for age **(problem 3)**
Appendix D : Analyses of infarct-like lesions by smoking history, age, and stroke **(problem 4)**

**Problem 5** asks you to consider the impact of multiple comparisons on the generalizability of statistical inference.

1. **Appendix B** provides results from three unadjusted analyses of the prevalence of infarct-like lesions by age.

    a. For each of the three models, can the model be used to **test for an association between prevalence of infarct-like lesions and age**? If so, can you tell what it is from the output? If so, provide the P value. If not, explain what further information you would need from the model. (Circle "yes" or "no" and either fill in the blank or supply just enough of an answer to let me know that you know what you would do.)

**Model B1:** Can be used to test for an association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

**Model B2:** Can be used to test for an association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

**Model B3:** Can be used to test for an association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

    b. For each of the three models, can the model be used to **test whether any association between prevalence of infarct-like lesions and age is nonlinear**? If so, can you tell what it is from the output? If so, provide the P value. If not, explain what further information you would need from the model. (Circle "yes" or "no" and either fill in the blank or supply just enough of an answer to let me know that you know what you would do.)

**Model B1:** Can be used to test for a nonlinear association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

**Model B2:** Can be used to test for a nonlinear association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

**Model B3:** Can be used to test for a nonlinear association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

c. For each of the three models, can the model be used to **test whether any association between prevalence of infarct-like lesions and age is U-shaped**? If so, can you tell what it is from the output? If so, provide the P value. If not, explain what further information you would need from the model. (Circle "yes" or "no" and either fill in the blank or supply just enough of an answer to let me know that you know what you would do.)

**Model B1:** Can be used to test for a U-shaped association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

**Model B2:** Can be used to test for a U-shaped association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

**Model B3:** Can be used to test for a U-shaped association?: **Yes** **No**

If so, is the relevant P value is available on output?: **Yes (P:_____)** **No**

If it can be used, but additional output is needed, what analysis do you need?

2. **Appendix C** provides results from three unadjusted analyses of the prevalence of infarct-like lesions by history of smoking behavior.

    a. For each of the three models, provide a scientifically useful interpretation of the estimated ***intercept***.

**Model C1:**

**Model C2:**

**Model C3:**

    b. For each of the three models, provide a scientifically useful interpretation of the estimated ***slope for pack years of smoking***.

**Model C1:**

**Model C2:**

**Model C3:**

    c. For each of the three models, provide a scientifically useful interpretation of the estimated ***slope for years since quitting.***.

**Model C1:**

**Model C2:**

**Model C3:**

d. What was the importance of using the "robust" standard errors in models C1 and C2? Should they also have been used in model C3?

e. Briefly discuss the relative advantages and disadvantages of the three regression models for exploring associations (both unadjusted and adjusted for confounding) in this setting**.** How similar are the results presented in Appendix B, and what issues might you worry about as you explore additional models?

3. Is the association between prevalence of infarct-like lesions and smoking history confounded by age? Briefly explain your reasoning. (**Appendix D** provides results of logistic regression analyses adjusted for age.)

4. **Appendix E** contains results of analyses of prevalence of infarct-like lesions by prior smoking history, age, and stroke.
    a. Provide scientifically useful interpretations for each of the regression parameters in the regression model. (Extreme brevity is OK)

b. What parameters would you test in order to answer the question of whether prior history of stroke modifies the association between infarct-like lesions and smoking?

c. What parameters would you test in order to answer the question of an association between infarct-like lesions and smoking?

d. How might the analysis results from the model in Appendix E that pertain to the group of patients without prior history of stroke compare to an analysis that restricted only to that group?

5. Suppose that historically approximately 10% of the hypotheses generated from exploratory analyses eventually pan out to be true. Suppose that we design confirmatory studies with a type 1 error of 0.05 and 90% statistical power. Explain the impact that performing multiple comparisons might have on the generalizability of the eventual analysis results. Consider both the cases of failure to adjust for inflation of the type 1 error the setting in which appropriate adjustment is made. (It is perfectly acceptable to illustrate the issues with hypothetical results. Just make clear those hypothetical assumptions.)

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

These data come from a cross-sectional study of cerebral MRI findings in a population based sample of 3,775 generally healthy older adults in four regions of the U.S. We are particularly interested in associations between smoking and the prevalence of infarct-like lesions on MRI. This exam considers the following variables (all measured at time of study enrolment).

| | |
|---|---|
| ***age*** | age of the subject in years |
| ***packyrs*** | pack year history (packs per day times years smoking) for the subject (never smokers = 0) |
| ***yrsquit*** | years since quitting smoking for the subject (never smokers = 0 and current smokers = 0) |
| ***stroke*** | categorical variable indicating prior history of cerebrovascular disease (0= none, 1= TIAs only, 2= clinical stroke) |
| ***infarcts*** | indicator of infarct-like lesions found on MRI (0= no, 1= yes) |

**Univariate statistics both overall and within groups defined by presence of infarct-like lesions:**

```
. tabstat age packyrs yrsquit stroke, by(infarcts) stat(n mean sd min q max)
```

| infarcts | vrbl | N | mean | sd | min | p25 | p50 | p75 | max |
|---|---|---|---|---|---|---|---|---|---|
| 0 | age | 2529 | 74.67 | 4.96 | 65 | 71 | 74 | 78 | 94 |
| | packyrs | 2439 | 15.69 | 23.99 | 0 | 0 | 0.50 | 26.0 | 204 |
| | yrsquit | 2529 | 8.67 | 13.83 | 0 | 0 | 0 | 15.0 | 83 |
| | stroke | 2529 | 0.058 | 0.313 | 0 | 0 | 0 | 0 | 2 |
| 1 | age | 1246 | 75.93 | 5.38 | 65 | 72 | 75 | 80 | 97 |
| | packyrs | 1209 | 21.19 | 29.21 | 0 | 0 | 6.75 | 36.0 | 240 |
| | yrsquit | 1246 | 7.67 | 13.18 | 0 | 0 | 0 | 12.0 | 74 |
| | stroke | 1246 | 0.295 | 0.684 | 0 | 0 | 0 | 0 | 2 |
| Total | age | 3775 | 75.09 | 5.14 | 65 | 71 | 74 | 78 | 97 |
| | packyrs | 3648 | 17.51 | 25.96 | 0 | 0 | 1.87 | 29.2 | 240 |
| | yrsquit | 3775 | 8.34 | 13.63 | 0 | 0 | 0 | 15.0 | 83 |
| | stroke | 3775 | 0.136 | 0.482 | 0 | 0 | 0 | 0 | 2 |

**Correlations among selected variables:**

```
. corr packyrs yrsquit age
(obs=3648)
```

| | packyrs | yrsquit | age |
|---|---|---|---|
| packyrs | 1.0000 | | |
| yrsquit | 0.0990 | 1.0000 | |
| age | -0.0975 | 0.0634 | 1.0000 |

**APPENDIX B**: **Selected analyses of prevalence of infarct-like lesions by age.**

**Creation of variables used in some analyses: squared age and linear splines with two knots**

```
. g agesqr= age^2
. mkspline sage65 72 sage72 80 sage80 = age
```

## Model B1: Logistic regression of *infarcts* on *age*

```
. logistic infarcts age
```

```
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(1)      =      49.26
                                                  Prob > chi2     =     0.0000
Log likelihood = -2369.5705                       Pseudo R2       =     0.0103

------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.047906   .0069919     7.01   0.000     1.034291    1.061699
------------------------------------------------------------------------------
```

## Model B2: Logistic regression of *infarcts* on *age* and *agesqr*

```
. logistic infarcts age agesqr
```

```
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(2)      =      49.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -2369.3959                       Pseudo R2       =     0.0104

------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.147666   .1769108     0.89   0.372     .8484068    1.552483
      agesqr |   .9994127   .0009943    -0.59   0.555     .9974659    1.001363
------------------------------------------------------------------------------
```

## Model B3: Logistic regression of *infarcts* on linear splines

```
. logistic infarcts sage*
```

```
Logistic regression                               Number of obs   =       3775
                                                  LR chi2(3)      =      49.91
                                                  Prob > chi2     =     0.0000
Log likelihood = -2369.2477                       Pseudo R2       =     0.0104

------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sage65 |   1.064267   .0365936     1.81   0.070     .9949078    1.138461
      sage72 |   1.050369   .0152387     3.39   0.001     1.020922    1.080665
      sage80 |   1.035128   .0213013     1.68   0.093     .9942086    1.077731
------------------------------------------------------------------------------
```

**APPENDIX C: Unadjusted analyses of prevalence of infarct-like lesions by pack year history and years since quitting smoking.**

## Model C1: Linear regression of *infarcts* on *packyrs, yrsquit*

```
. regress infarcts packyrs yrsquit, robust

Linear regression                               Number of obs =    3648
                                                F(  2,  3645) =   21.46
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0122
                                                Root MSE      =  .46803
------------------------------------------------------------------------------
             |               Robust
    infarcts |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     packyrs |   .0018941   .0003086     6.14   0.000      .001289    .0024991
     yrsquit |   -.001678    .000576    -2.91   0.004    -.0028073   -.0005487
       _cons |    .312111   .0102084    30.57   0.000     .2920963    .3321256
------------------------------------------------------------------------------
```

## Model C2: Poisson regression of *infarcts* on *packyrs, yrsquit*

```
. poisson infarcts packyrs yrsquit, robust

Poisson regression                              Number of obs   =       3648
                                                Wald chi2(2)    =      55.65
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -2530.5488               Pseudo R2       =     0.0054
------------------------------------------------------------------------------
             |               Robust
    infarcts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     packyrs |   .0048917   .0006839     7.15   0.000     .0035513    .0062321
     yrsquit |  -.0051946   .0019799    -2.62   0.009    -.0090751   -.0013141
       _cons |  -1.157517   .0313646   -36.91   0.000    -1.218991   -1.096044
------------------------------------------------------------------------------
```

## Model C3: Logistic regression of *infarcts* on *packyrs, yrsquit*

```
. logit infarcts packyrs yrsquit

Logistic regression                             Number of obs   =       3648
                                                LR chi2(2)      =      43.39
                                                Prob > chi2     =     0.0000
Log likelihood = -2295.4261                     Pseudo R2       =     0.0094
------------------------------------------------------------------------------
    infarcts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     packyrs |   .0082182    .001333     6.17   0.000     .0056055    .0108309
     yrsquit |   -.007836   .0027611    -2.84   0.005    -.0132477   -.0024243
       _cons |  -.7882922    .047226   -16.69   0.000    -.8808536   -.6957309
------------------------------------------------------------------------------
```

**<u>APPENDIX D</u>: Age adjusted analyses of the association between prevalence of infarct-like lesions and smoking history.**

```
. logit infarcts packyrs yrsquit age

Logistic regression                                     Number of obs   =       3648
                                                        LR chi2(3)      =     105.68
                                                        Prob > chi2     =     0.0000
Log likelihood = -2264.2824                             Pseudo R2       =     0.0228

------------------------------------------------------------------------------
    infarcts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     packyrs |   .0095282   .0013604     7.00   0.000     .0068618    .0121946
     yrsquit |  -.0095575   .0027866    -3.43   0.001    -.0150192   -.0040957
         age |   .0547071   .0069528     7.87   0.000     .0410799    .0683343
       _cons |  -4.916088   .5284127    -9.30   0.000    -5.951758   -3.880419
------------------------------------------------------------------------------
```

```
. logistic infarcts packyrs yrsquit age

Logistic regression                                     Number of obs   =       3648
                                                        LR chi2(3)      =     105.68
                                                        Prob > chi2     =     0.0000
Log likelihood = -2264.2824                             Pseudo R2       =     0.0228

------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     packyrs |   1.009574   .0013735     7.00   0.000     1.006885    1.012269
     yrsquit |   .9904881   .0027601    -3.43   0.001      .985093    .9959126
         age |   1.056231   .0073437     7.87   0.000     1.041935    1.070723
------------------------------------------------------------------------------
```

**APPENDIX E: Analyses of the association between prevalence of infarct-like lesions and smoking history, age, and prior history of stroke.**

```
. logit infarcts i.stroke##c.packyrs i.stroke##c.yrsquit age
Logistic regression                               Number of obs   =       3648
                                                  LR chi2(9)      =     268.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -2182.7282                       Pseudo R2       =     0.0580

------------------------------------------------------------------------------
    infarcts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      stroke |
          1  |   .8295187   .3332585     2.49   0.013      .176344    1.482693
          2  |   1.837737   .2258669     8.14   0.000     1.395046    2.280428
             |
     packyrs |   .0085575   .0014599     5.86   0.000     .0056962    .0114188
             |
      stroke#|
   c.packyrs |
          1  |   .0066201   .0101546     0.65   0.514    -.0132826    .0265228
          2  |   .0054338   .0069287     0.78   0.433    -.0081461    .0190138
             |
     yrsquit |  -.0097968   .0029852    -3.28   0.001    -.0156478   -.0039459
             |
      stroke#|
   c.yrsquit |
          1  |   .0100385   .0200801     0.50   0.617    -.0293178    .0493949
          2  |  -.0082753   .0117769    -0.70   0.482    -.0313576    .0148069
             |
         age |   .0517036   .0071459     7.24   0.000     .0376979    .0657094
       _cons |  -4.805626   .5429103    -8.85   0.000    -5.869711   -3.741541
------------------------------------------------------------------------------

. logistic infarcts i.stroke##c.packyrs i.stroke##c.yrsquit age
Logistic regression                               Number of obs   =       3648
                                                  LR chi2(9)      =     268.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -2182.7282                       Pseudo R2       =     0.0580

------------------------------------------------------------------------------
    infarcts | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      stroke |
          1  |   2.292215   .7639002     2.49   0.013     1.192848    4.404793
          2  |   6.282307   1.418965     8.14   0.000     4.035162    9.780868
             |
     packyrs |   1.008594   .0014724     5.86   0.000     1.005712    1.011484
             |
      stroke#|
   c.packyrs |
          1  |   1.006642   .0102221     0.65   0.514     .9868053    1.026878
          2  |   1.005449   .0069664     0.78   0.433      .991887    1.019196
             |
     yrsquit |    .990251   .0029561    -3.28   0.001      .984474    .9960619
             |
      stroke#|
   c.yrsquit |
          1  |   1.010089   .0202827     0.50   0.617     .9711078    1.050635
          2  |   .9917588   .0116798    -0.70   0.482      .969129    1.014917
             |
         age |   1.053064   .0075251     7.24   0.000     1.038417    1.067916
------------------------------------------------------------------------------
```