Stata tips #6 (STATA session#6)
Epidemiologic Methods
Adapted from notes of previous TAs

Logistic regression in Stata

Unique points regarding logistic regression:

- Response variable is binary
  - Always code outcome variable as 01 and remember which is 1!!!
  - e.g. death01 (where 0=alive, 1=dead)
- Results are interpreted in terms of log-odds or odds ratios

Two main Stata commands:

- *logit* – reports results as coefficients; can add the *or* option to get odds ratios
- *logistic* – reports results as odds ratios

The prediction equation is:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Typical interpretations in simple logistic regression:

- When a coefficient is reported: For a one-unit increase in X_1, the log-odds of the response variable increase by β_1, holding all other covariates constant.
- When an odds ratio is reported: The odds of the response increases by (OR) for a one-unit increase in X_1, holding all other covariates constant.

We will use death as our outcome, and consider associations with age, BMI, and diabetes (see variable creation below). We will use the data in wide format.

**BMI categories

```
gen bmicat=2 if bmi>=18.5 & bmi<25
replace bmicat=1 if bmi<18.5 & bmi~=.
replace bmicat=3 if bmi>=25 & bmi<30
replace bmicat=4 if bmi>=30 & bmi<35
replace bmicat=5 if bmi>=35 & bmi~=.
```

**Cholesterol categories

```
gen highchol=1 if totchol>=200 & totchol~=.
replace highchol=0 if totchol<200 & totchol~=.
```

*Age categories

```
gen agecat=.
replace agecat=0 if age<40
```

```
replace agecat=1 if age>39
replace agecat=2 if age>50
replace agecat=3 if age>60
```

*Reshape

```
reshape wide totchol age sysbp diabp cursmoke cigpday bmi
diabetes bpmeds heartrte glucose educ prevchd prevap prevmi
prevstrk prevhyp time hdlc ldlc death angina hospmi mi_fchd
anychd stroke cvd hyperten timeap timemi timemifc timechd
timestrk timecvd timedth timehyp bmicat highchol agecat,
i(randid) j(period)
```

```
logit death1 age1
```

```
Iteration 0:   log likelihood = -2869.6004
Iteration 1:   log likelihood =  -2460.167
Iteration 2:   log likelihood = -2453.1547
Iteration 3:   log likelihood = -2453.1496
Iteration 4:   log likelihood = -2453.1496

Logistic regression                                     Number of obs   =       4434
                                                        LR chi2(1)      =     832.90
                                                        Prob > chi2     =     0.0000
Log likelihood = -2453.1496                             Pseudo R2       =     0.1451

------------------------------------------------------------------------------
      death1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |   .1148415   .0043927    26.14   0.000     .1062321     .123451
       _cons |  -6.484358   .2309887   -28.07   0.000    -6.937088   -6.031629
------------------------------------------------------------------------------
```

The "iteration" portion of the output represents Stata's maximum likelihood iteration computations, which converge at the point of maximum log likelihood. This portion of the output is useful for checking model fitting (if taking many iterations, it indicates an issue for the model fit). It can be omitted with the "nolog" option to save space:

```
logit death1 age1, nolog
```

```
Logistic regression                                     Number of obs   =       4434
                                                        LR chi2(1)      =     832.90
                                                        Prob > chi2     =     0.0000
Log likelihood = -2453.1496                             Pseudo R2       =     0.1451

------------------------------------------------------------------------------
      death1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |   .1148415   .0043927    26.14   0.000     .1062321     .123451
       _cons |  -6.484358   .2309887   -28.07   0.000    -6.937088   -6.031629
------------------------------------------------------------------------------
```

The interpretation here from "Coef." is that each one year increase in age is associated with a change of 0.11 in the log-odds of death. This finding is statistically significant ($p < 0.001$, 95% CI = 0.106-0.123). Remember that the Z value is just the coefficient divided by the standard error. Also note that the R2 is a Pseudo R2 and does not have the

same interpretation as in linear regression (which is?). The change in the log odds of death is a mouthful to say and hard to wrap your head around. Let's use the "or" option:

```
. logit death1 age1, or nolog

Logistic regression                                     Number of obs     =      4,434
                                                        LR chi2(1)        =     832.90
                                                        Prob > chi2       =     0.0000
Log likelihood = -2453.1496                             Pseudo R2         =     0.1451

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |   1.121696   .0049272    26.14   0.000      1.11208    1.131395
       _cons |   .0015271   .0003528   -28.07   0.000     .0009711    .0024016
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Now, we get our results above as an odds ratio instead of a coefficient. The interpretation here is that for each one-year increase in age, the odds of death increases by 12%, and again this result is statistically significant.

You can also use the *logistic* command to get the same output (without the "or" option):

```
. logistic death1 age1, nolog

Logistic regression                                     Number of obs     =      4,434
                                                        LR chi2(1)        =     832.90
                                                        Prob > chi2       =     0.0000
Log likelihood = -2453.1496                             Pseudo R2         =     0.1451

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age1 |   1.121696   .0049272    26.14   0.000      1.11208    1.131395
       _cons |   .0015271   .0003528   -28.07   0.000     .0009711    .0024016
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Note: What LR and chi2 are telling you

- LR chi2( ) - This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model.
- Prob > chi2 - This is the probability of getting a LR test statistic as extreme as, or more so, than the observed under the null hypothesis; the null hypothesis is that all of the regression coefficients in the model are equal to zero. In other words, this is the probability of obtaining the chi-square statistic if there is in fact no effect of the predictor variables.

Confounder control with continuous covariates

In most epidemiological studies, you are interested in the relationship between a predictor and an outcome, but you need to control for other covariates (confounders). When these covariates are nominal (e.g. religious affiliation) you have limited options, and your categories are determined by the detail of the data you collected and decisions on how to potentially lump categories together. When your confounder covariates are continuous, however, you have an almost unlimited array of options. The decisions you make regarding how to code your continuous confounder variables can have a major impact on the validity of your study results.

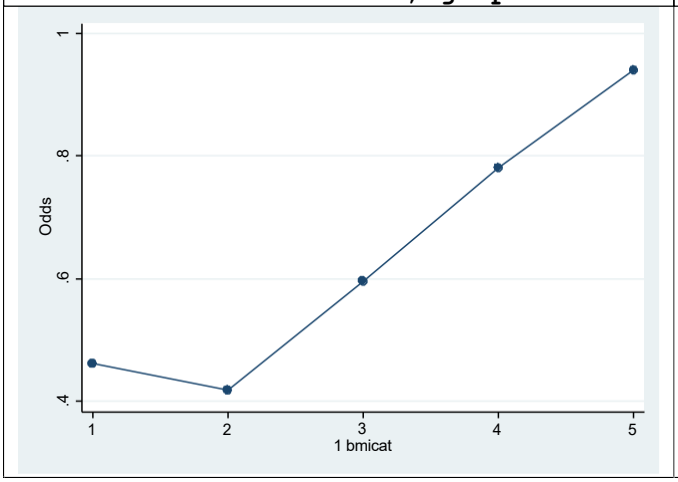
There are many options to choose from and each has their pros and cons:

1. Keep the continuous variable continuous
    a. Great when the underlying "true" relationship is nearly linear, but can perform poorly when it is not
2. Dichotomize the variable
    a. Usually not a good idea! (We do this in class occasionally for simplicity and to practice interpretations and interactions, or if dichotomization has clinical significance.)
    b. Often results in residual confounding (can you think of a continuous variable that has a true cut-point delineating 0% and 100% risk?)
3. Divide the variable into multiple categories
    a. Usually need 4 or so categories to control for confounding
    b. Good if there is prior literature to guide where to categorize
    c. Works pretty well if relationship is non-linear
    d. However, often unclear where to make cut-points (and they might be different in different studies) and still unnatural (risk is not usually a step-function)
4. Keep continuous but change functional form
    a. Often provides best confounder control when done correctly and is more natural
    b. More difficult to interpret and some options complex
    c. Options:
        i. Polynomials (e.g. add a square term)
        ii. Fractional polynomials (e.g. mfp in Stata)
        iii. Splines (e.g. mvrs in Stata)

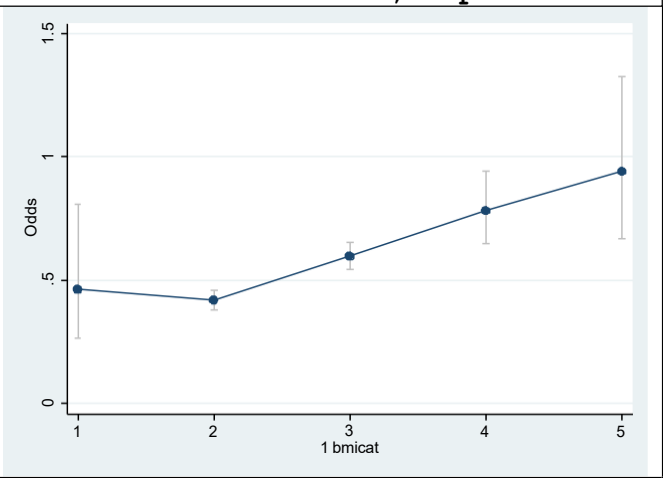
We will focus on the first three options. Remember, because this is confounder control, we are not as concerned with p-values when determining whether to keep a variable in the equation. What we care about is how the variable of interest changes with and without the confounder.

Let's look at some examples. Say we were interested in whether diabetes is associated with death, controlling for age, BMI, cholesterol level, and smoking. First, let's think about how we might want to control for BMI as a confounder. Because it's a continuous covariate, and prior literature has strongly suggested cut-points, let's see how they relate to our outcome of interest.

`tabodds death1 bmicat1, graph`

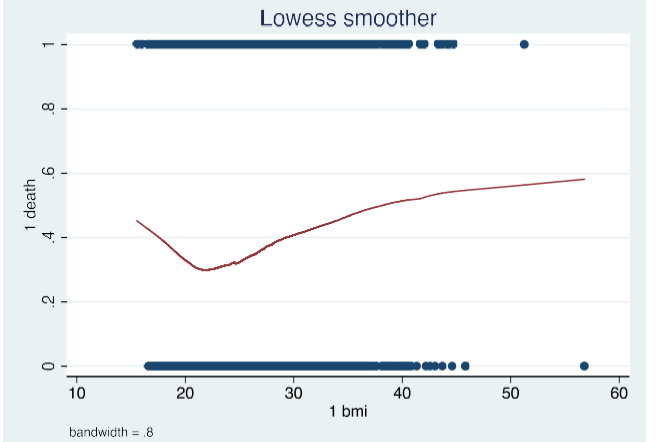


`tabodds death1 bmicat1, ciplot`

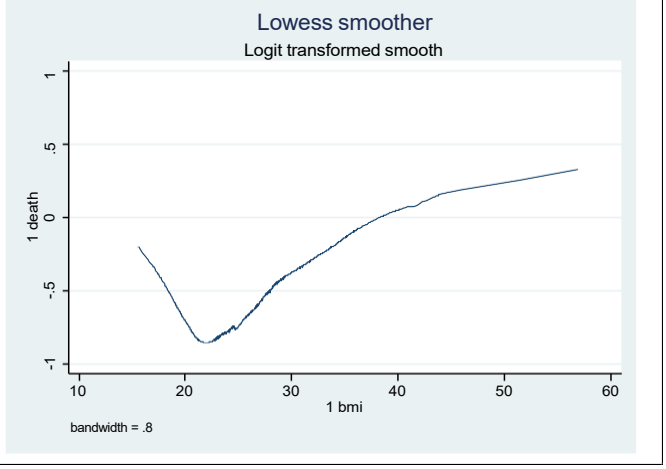


So what do you think? Should we try nominal (categorical) coding? Ordinal coding? What does the relationship look like with death versus continuous BMI?

`lowess death1 bmi1`



`lowess death1 bmi1, logit`



<u>Note</u>: no need to perform `lowess` for your homework for two reasons; it will take a long time to run in the bigsample dataset and you sometimes need to play with the options to make it look right. More importantly, lowess performs a linear model, while death/alive is a binary outcome, so the shape of relation might be misleading.

The lowess smoother suggests that the categories we picked capture the "true" relationship and so it seems like categorical coding would be better than continuous or ordinal coding. Let's see what the regression tells us:

**Linear

```
. logistic death1 bmi1, nolog

Logistic regression                                     Number of obs     =      4,415
                                                        LR chi2(1)        =      44.28
                                                        Prob > chi2       =     0.0000
Log likelihood = -2831.1941                             Pseudo R2         =     0.0078

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmi1 |   1.052252   .0080871     6.63   0.000      1.03652    1.068222
       _cons |   .1423284   .0288676    -9.61   0.000     .0956418    .2118046
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

**Ordinal

```
. logistic death1 bmicat1, nolog

Logistic regression                                     Number of obs     =      4,415
                                                        LR chi2(1)        =      53.33
                                                        Prob > chi2       =     0.0000
Log likelihood = -2826.6709                             Pseudo R2         =     0.0093

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bmicat1 |   1.338571   .0535494     7.29   0.000     1.237625     1.44775
       _cons |   .2415579   .0275975   -12.43   0.000      .193096    .3021823
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

**Categorical

```
. logistic death1 i.bmicat1, nolog

Logistic regression                                     Number of obs     =      4,415
                                                        LR chi2(4)        =      55.96
                                                        Prob > chi2       =     0.0000
Log likelihood = -2825.3577                             Pseudo R2         =     0.0098

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bmicat1 |
          2  |   .9063492   .2621849    -0.34   0.734     .5141189    1.597819
          3  |   1.291378   .3731905     0.88   0.376     .7329393    2.275302
          4  |   1.691899   .5083662     1.75   0.080     .9388869    3.048847
          5  |   2.037313   .6818002     2.13   0.033     1.057299    3.925708
             |
       _cons |   .4615385   .1315154    -2.71   0.007     .2640328    .8067851
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

So, what do you think? It looks like the ordinal is better than the continuous, based on the chi-squared with the same degrees of freedom (1). If you were to test the categorical coding against the ordinal (e.g. by AIC), I bet they would be about the same (or ordinal might even be better) because there are so few people in the underweight group.

However, because we have plenty of degrees of freedom to spare in this huge dataset, and the relationship does look J-shaped (and this has been shown in the literature), I would pick categorical coding here.

```
. estimate stats s1 s2 s3

Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC         BIC
-------------+----------------------------------------------------------------------------------
          s1 |      4,415  -2853.336  -2825.358       5   5660.715    5692.679
          s2 |      4,415  -2853.336  -2826.671       2   5657.342    5670.127
          s3 |      4,415  -2853.336  -2831.194       2   5666.388    5679.174
------------------------------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

The next question to address: is BMI a confounder in the relationship between diabetes and death?

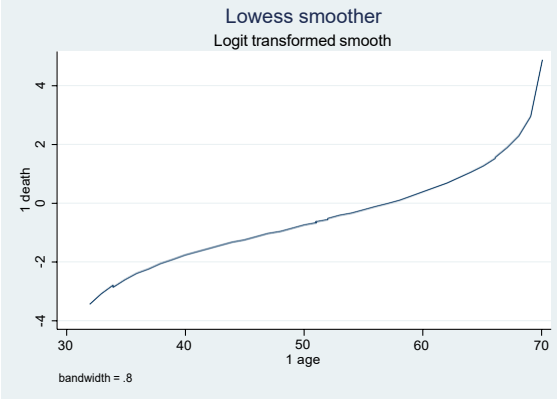
```
. logistic death1 diabetes1, nolog

Logistic regression                                     Number of obs     =      4,434
                                                        LR chi2(1)        =      95.10
                                                        Prob > chi2       =     0.0000
Log likelihood = -2822.0517                             Pseudo R2         =     0.0166

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--
   diabetes1 |   6.831451   1.507758     8.71   0.000     4.432469    10.52883
       _cons |   .5096255   .0164099   -20.93   0.000     .4784566    .5428248
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. logistic death1 diabetes1 bmi1, nolog

Logistic regression                                     Number of obs     =      4,415
                                                        LR chi2(2)        =      127.20
                                                        Prob > chi2       =     0.0000
Log likelihood = -2789.7359                             Pseudo R2         =     0.0223

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   diabetes1 |     6.2013   1.379933     8.20   0.000      4.00932    9.591682
        bmi1 |   1.046942   .0081662     5.88   0.000     1.031058     1.06307
       _cons |   .1544863   .0317294    -9.09   0.000     .1032915     .231055
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

```
. . logistic death1 diabetes1 bmicat1, nolog

Logistic regression                                     Number of obs     =      4,415
                                                        LR chi2(2)        =     136.46
                                                        Prob > chi2       =     0.0000
Log likelihood = -2785.1053                             Pseudo R2         =     0.0239

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   diabetes1 |   6.218682   1.384335     8.21   0.000     4.019881    9.620186
     bmicat1 |   1.308579   .0530801     6.63   0.000     1.208572    1.416861
       _cons |   .2445182   .0282441   -12.19   0.000     .1949799    .3066428
------------------------------------------------------------------------------

. logistic death1 diabetes1 i.bmicat1, nolog

Logistic regression                                     Number of obs     =      4,415
                                                        LR chi2(5)        =     138.72
                                                        Prob > chi2       =     0.0000
Log likelihood = -2783.9755                             Pseudo R2         =     0.0243

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   diabetes1 |   6.194699   1.379171     8.19   0.000     4.004156    9.583615
             |
     bmicat1 |
          2  |   .9760088   .2884408    -0.08   0.935     .5468867    1.741847
          3  |   1.362453   .4022485     1.05   0.295     .7638606    2.430127
          4  |   1.749504   .5367493     1.82   0.068     .9588829     3.19201
          5  |    1.96326   .6720348     1.97   0.049     1.003706    3.840155
             |
       _cons |   .4150947   .1209397    -3.02   0.003     .2345016    .7347652
------------------------------------------------------------------------------
```

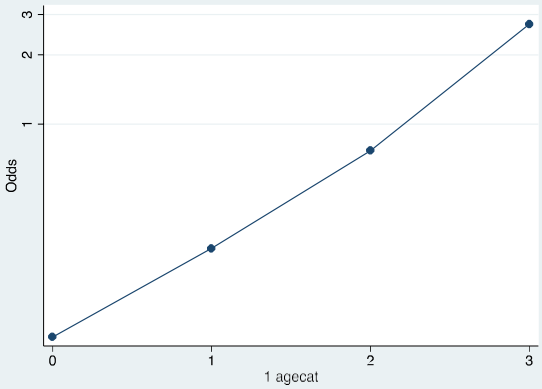
So the odds ratio changes from 6.83 (crude) to 6.19 (adjusted). Is there evidence of confounding? Did it matter here what functional form BMI took on? How would you interpret the results of the study?

How would we model age?

lowess death1 age1, logit



tabodds death1 agecat1, graph yscale(log)




```
. logistic death1 diabetes1 age1, nolog

Logistic regression                                     Number of obs     =      4,434
                                                        LR chi2(2)        =     887.86
                                                        Prob > chi2       =     0.0000
Log likelihood = -2425.6691                             Pseudo R2         =     0.1547

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   diabetes1 |   4.962262   1.167647     6.81   0.000     3.128869    7.869951
        age1 |   1.119846   .0049492    25.61   0.000     1.110187    1.129588
       _cons |      .00159   .0003695   -27.73   0.000     .0010083    .0025074
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

```
. logistic death1 diabetes1 agecat1, nolog

Logistic regression                                     Number of obs     =      4,434
                                                        LR chi2(2)        =     809.77
                                                        Prob > chi2       =     0.0000
Log likelihood = -2464.7146                             Pseudo R2         =     0.1411

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   diabetes1 |   5.169815   1.216486     6.98   0.000     3.259737    8.199123
     agecat1 |   2.840893   .1217144    24.37   0.000     2.612079    3.089751
       _cons |   .0974303   .0077855   -29.14   0.000     .0833059    .1139496
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

```
. logistic death1 diabetes1 i.agecat1, nolog

Logistic regression                                     Number of obs     =      4,434
                                                        LR chi2(4)        =     815.72
                                                        Prob > chi2       =     0.0000
Log likelihood = -2461.7421                             Pseudo R2         =     0.1421

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   diabetes1 |   5.204498   1.223799     7.01   0.000     3.282654    8.251495
             |
     agecat1 |
          1  |   2.384156   .3554108     5.83   0.000       1.7801    3.193191
          2  |   6.252884   .9283854    12.35   0.000      4.67412    8.364904
          3  |   21.95231   3.594858    18.86   0.000     15.92537    30.26015
             |
       _cons |   .1161736   .0160299   -15.60   0.000     .0886455    .1522503
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Which method do you prefer here? The beta coefficient of diabetes is lowest for the linear form of age as confounding and the model chi-squared value is also highest which suggests that the linear relationship fits the data really well so I would go with this one. How would you interpret the results?

Effect Modification in Logistic Regression

How would we answer the question: is smoking status an effect modifier in the relationship between diabetes and death?

```
. logit death1 i.diabetes1##i.cursmoke1, nolog

Logistic regression                                     Number of obs     =      4,434
                                                        LR chi2(3)        =     101.86
                                                        Prob > chi2       =     0.0000
Log likelihood = -2818.6684                             Pseudo R2         =     0.0177

------------------------------------------------------------------------------
             death1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------
        1.diabetes1 |   2.199705   .2905199     7.57   0.000     1.630296    2.769113
        1.cursmoke1 |   .1489921    .064448     2.31   0.021     .0226764    .2753079
                    |
diabetes1#cursmoke1 |
              1 1   |  -.6835344   .4506793    -1.52   0.129     -1.56685    .1997809
                    |
              _cons |  -.7488719   .0459366   -16.30   0.000     -.838906   -.6588378
------------------------------------------------------------------------------
```

Is smoking an effect modifier? How do we interpret the output?
What are the smoking status-specific odds ratios for the association between diabetes and death?

Non-smokers:

lincom 1.diabetes1

```
 ( 1)  [death1]1.diabetes1 = 0

------------------------------------------------------------------------------
      death1 |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.199705   .2905199     7.57   0.000     1.630296    2.769113
------------------------------------------------------------------------------
```

lincom 1.diabetes1, or

```
 ( 1)  [death1]1.diabetes1 = 0

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   9.022348   2.621172     7.57   0.000     5.105386    15.94449
------------------------------------------------------------------------------
```

Smokers:

lincom 1.diabetes1 + 1.diabetes1#1.cursmoke1, or

```
 ( 1)  [death1]1.diabetes1 + [death1]1.diabetes1#1.cursmoke1 = 0

------------------------------------------------------------------------------
      death1 | Odds Ratio   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   4.554748   1.569308     4.40   0.000     2.318402    8.948291
------------------------------------------------------------------------------
```

Note that the interaction term in the model (on the odds ratio scale) is (4.55/ 9.02) = 0.50