12 Odds Ratios for Multi-level Factors; Examples

The Framingham Study

The Framingham study was a prospective (follow-up, cohort) study of the occurrence of coronary heart disease (CHD) in Framingham, Mass. The study involved 2187 men and 2669 women aged between 30 and 62. More details on the study are given as a supplement to the lecture. Variables and values of the variables are as follows:

Variable Name	Codes
Gender	0 = Female, 1 = male
Age Group	0 is 30-49, 1 is 50-62
SCL (Serum Cholesterol)	1 is < 190, 2 is 190-219, 3 is 220-249, 4 is 250+
CHD (Coronary Heart Disease)	1 is Yes, 0 is No
Freq	Count

I will consider a simple analysis of the association between serum cholesterol level (SCL) at the start of the study and whether a subject had, or developed CHD, during the 12 year follow-up period. A table with **Stata** analysis of counts relating CHD to SCL is given below.

```
. tabulate chd scl [fw=frequency],chi2 lrchi2 exp col
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| column percentage  |
+--------------------+

           |                        SCL
       CHD |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         0 |     1,022      1,203      1,119      1,125 |     4,469
           |     978.3    1,169.7    1,127.4    1,193.6 |   4,469.0
           |     96.14      94.65      91.35      86.74 |     92.03
-----------+--------------------------------------------+----------
         1 |        41         68        106        172 |       387
           |      84.7      101.3       97.6      103.4 |     387.0
           |      3.86       5.35       8.65      13.26 |      7.97
-----------+--------------------------------------------+----------
     Total |     1,063      1,271      1,225      1,297 |     4,856
           |   1,063.0    1,271.0    1,225.0    1,297.0 |   4,856.0
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =  86.7040   Pr = 0.000
 likelihood-ratio chi2(3) =  85.8644   Pr = 0.000
```

The Pearson χ^2 statistic, which can be viewed as testing that the probability of developing CHD is independent of SCL, is highly significant (p-value < .001). Clearly observed counts of CHD are below expected counts for this hypothesis with low SCL, and above with high SCL, so it looks like CHD increases as SCL increases.

Let us do a closer look at the data for CHD vs. SCL using odds ratios. There are a lot of possible ways to do this. Since SCL categories are ordered, many analysts would compare SCL level 2 to 1, then 3 to 2, then 4 to 3. It is a little more conventional (and slightly more direct to implement in **Stata**) to consider all OR relative to a fixed baseline SCL category, say SCL < 190 (Cat. 1).

	SCL		
CHD	2	1	
Y	68	41	
N	1203	1022	$\widehat{OR}(2\text{vs.}1) = \frac{68 \cdot 1022}{41 \cdot 1203} = 1.409$

	3	1	
Y	106	41	
N	1119	1022	$\widehat{OR}(3\text{vs.}1) = \frac{106 \cdot 1022}{41 \cdot 1119} = 2.361$

	4	1	
Y	172	41	
N	1125	1022	$\widehat{OR}(4\text{vs.}1) = \frac{172 \cdot 1022}{41 \cdot 1125} = 3.811$

Any OR may be computed from this set of OR's. For example,

	SCL		
CHD	4	2	
Y	172	68	
N	1125	1203	$\widehat{OR}(4\text{vs.}2) = \frac{172 \cdot 1203}{1125 \cdot 68} = 2.705 = \frac{3.811}{1.409} = \frac{\widehat{OR}(4\text{vs.}1)}{\widehat{OR}(2\text{vs.}1)}$

Think of this relationship as $\frac{4}{2} = \frac{4/1}{2/1}$. An important point to recognize is that the effect of SCL on CHD can be captured through 3 effects (ORs), which is #SC levels - 1.

To get these ORs directly from **Stata**, we need to use `xi`. Actually, there are other, better, options you can download and install, like `xi3` and `desmat`. Since `xi` is built-in and commonly used, we will stick with it but it does not allow higher order interaction terms in models, unlike `xi3` and `desmat`.

The code and output follow:

```
. xi:logistic chd i.scl [fweight=frequency]
 i.scl            _Iscl_1-4        (naturally coded;_Iscl_1 omitted)
Logistic regression                               Number of obs   =       4856
                                                  LR chi2(3)      =      85.86
                                                  Prob > chi2     =     0.0000
Log likelihood = -1307.1541                       Pseudo R2       =     0.0318
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iscl_2 |   1.408998   .2849726     1.70   0.090     .9478795    2.094438
     _Iscl_3 |   2.361255    .446123     4.55   0.000     1.630502    3.419514
     _Iscl_4 |   3.811035   .6825005     7.47   0.000     2.682905    5.413532
------------------------------------------------------------------------------
. xi:logistic chd i.scl [fweight=frequency],coef
------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iscl_2 |   .3428787    .202252     1.70   0.090    -.0535279    .7392852
     _Iscl_3 |   .8591931   .1889347     4.55   0.000     .4888878    1.229498
     _Iscl_4 |   1.337901   .1790853     7.47   0.000        .9869    1.688902
       _cons |  -3.215945   .1592756   -20.19   0.000    -3.528119    -2.90377
------------------------------------------------------------------------------
```

Remember what `xi` is doing. It creates indicator variables with the first omitted, so we are fitting a model for p = probability of CHD of

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_j \text{ \_Iscl\_j, where I\_scl\_j} = \begin{cases} 1 & \text{SCL} = j, j \geq 2 \\ 0 & \text{SCL} \neq j, j \geq 2 \\ 0 & j = 1 \text{ (i.e.naturally coded; \_Iscl\_1 omitted)} \end{cases}$$

and proceeding as in the last lecture, for $j > 1$,

$$\begin{aligned} \beta_j = & \ (\alpha + \beta_j) - \alpha \\ = & \ \log(\text{Odds for SCL} = j) - \log(\text{Odds for SCL} = 1) \\ = & \ \log(\text{OR(for SCL} = \text{j vs. SCL} = 1)) \end{aligned}$$

which yields the result that

$$e^{\beta_j} = \text{OR(for SCL} = \text{j vs. SCL} = 1) \qquad j > 1$$

with confidence intervals for ORs produced by exponentiating limits of confidence intervals for coefficients. The **Stata** output above gives us exactly the values of $\widehat{OR}$(2vs.1), $\widehat{OR}$(3vs.1), and $\widehat{OR}$(4vs.1) we calculated previously, along with confidence limits. We also saw that $\widehat{OR}(4\text{vs.}2) = \frac{\widehat{OR}(4\text{vs.}1)}{\widehat{OR}(2\text{vs.}1)} = \frac{3.811}{1.409} = 2.705$ but this does not produce a confidence interval for OR(4vs.2). In order to get full information about this OR, note that

$$\frac{\widehat{OR}(4\text{vs.}1)}{\widehat{OR}(2\text{vs.}1)} = \frac{e^{\beta_4}}{e^{\beta_2}} = e^{\beta_4 - \beta_2}$$

This looks like `lincom` should work, and it is exactly the solution.

```
. lincom _b[_Iscl_4] - _b[_Iscl_2]
 ( 1) - _Iscl_2 + _Iscl_4 = 0
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.704784   .4033665     6.67   0.000      2.01926    3.623039
------------------------------------------------------------------------------
```

`lincom` reports OR after `logistic`. If you actually want the difference in coefficients, you need to use the `logit` form of the command, and then `lincom` reports

```
. lincom _b[_Iscl_4] - _b[_Iscl_2]
 ( 1) - _Iscl_2 + _Iscl_4 = 0
------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9950222   .1491307     6.67   0.000     .7027313    1.287313
------------------------------------------------------------------------------
```

This section has shown you how to generate unadjusted ORs in **Stata**. In practice we would add confounding variables, such as Age and Sex, to the model and then evaluate adjusted ORs for the SCL levels. You will get to do this in lab.

Model Building

There are a variety of systematic approaches to logistic regression models. Automated methods such as the backward elimination approach described below are well suited for producing good predictive models. Systematic approaches such as those advocated in Kleinbaum's book on Logistic Regression focus more attention on understanding the complex interdependencies among the predictors, and their impact on odds ratios.

Backward Elimination

1. Identify predictors or factors for which an association with outcome is biologically plausible (based on literature, science, knowledge, etc.).
2. Identify biologically plausible interactions.
3. Fit the logistic model with all candidate effects identified in the first 2 steps.
4. Following the *hierarchy principle*, identify the least significant effect in the model, and sequentially eliminate the least significant effect until the step where the least significant effect is "too important" to omit.

- The *hierarchy principle* implies that a main effect in a model can only be considered for exclusion if the model does not contain an interaction involving the main effect.
- The impact of an effect is measured by a p-value for testing that the regression coefficient for the effect is zero.

$$\text{p} - \text{value} \begin{cases} \leq \alpha & \text{effect stays in model} \\ > \alpha & \text{effect is removed} \end{cases}$$

In backwards elimination it is not uncommon to set $\alpha = .10$ or .15 rather than $\alpha = .05$.

Example: UNM Trauma Data

Response	:	Death (1=yes, 0 = no)
Predictors	:	ISS
		Age
		RTS
		BP (0=Blunt, 1=Penetratin)

The surgeon who collected the data, Dr. Turner Osler, believes that all these effects are associated with the probability of death and that the three interactions involving BP (BP*ISS, BP*Age, BP*RTS) are plausible.

Steps

0. Fit full model

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5(\text{BP} * \text{ISS}) + \beta_6(\text{BP} * \text{RTS}) + \beta_7(\text{BP} * \text{Age})$$

where p=probability of death from injuries. **Stata** does not allow specification of interaction terms directly with `logit` or `logistic`, so we need to use `xi`.

```
. xi:logistic death iss i.bp rts age i.bp*iss i.bp*rts i.bp*age
i.bp              _Ibp_0-1          (naturally coded; _Ibp_0 omitted)
i.bp*iss          _IbpXiss_#        (coded as above)
i.bp*rts          _IbpXrts_#        (coded as above)
i.bp*age          _IbpXage_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(7)      =     937.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -443.88603                       Pseudo R2       =     0.5136
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070319   .0090198     8.06   0.000     1.052785    1.088144
         age |   1.047169   .0058718     8.22   0.000     1.035724    1.058741
  _IbpXiss_1 |   .9927199   .0161392    -0.45   0.653     .9615863    1.024861
```

```
         rts |   .4714732   .0285804   -12.40   0.000     .4186563    .5309533
  _IbpXrts_1 |   .7540543   .1137513    -1.87   0.061     .5610433    1.013465
      _Ibp_1 |   12.53281    14.3303     2.21   0.027       1.3328    117.8506
  _IbpXage_1 |   1.013542   .0148866     0.92   0.360     .9847811    1.043143
------------------------------------------------------------------------------
. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        16.84
                 Prob > chi2 =         0.0318
```

At this point I am not happy about the goodness-of-fit test. The objection raised earlier in class that age probably does not have a strictly linear effect may be coming back to bite us here. I hate to proceed with a full model that does not seem to fit well. I experimented with a couple of approaches to be more flexible with age. One was to create age groupings, the other was to fit an additional term that allowed curvature in age in the logit scale. The latter approach is more parsimonious and I liked the results more, although there was not a lot of difference (older patients are fit with greatly reduced odds of survival either way). The distribution of age already is skewed right in this data set, so instead of introducing a term for square of age I introduced a term for square root of age – the difference in logit fits being slight for older patients but substantial for very young ones. Now I fit the model above with this new age term introduced along with the interaction:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= \alpha+\beta_1\text{ISS}+\beta_2\text{BP}+\beta_3\text{RTS}+\beta_4\text{Age}+\beta_5\sqrt{\text{Age}}+\beta_6(\text{BP}*\text{ISS})\\ &\quad+\beta_7(\text{BP}*\text{RTS})+\beta_8(\text{BP}*\text{Age})+\beta_9(\text{BP}*\sqrt{\text{Age}})\end{aligned}$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*iss i.bp*rts i.bp*age i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*iss          _IbpXiss_#          (coded as above)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*age          _IbpXage_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(9)      =     944.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -440.23492                       Pseudo R2       =     0.5176
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.073518   .0092108     8.27   0.000     1.055616    1.091723
         age |   1.110938   .0256408     4.56   0.000     1.061802    1.162347
     agesqrt |   .4900429   .1303889    -2.68   0.007     .2909038    .8255033
  _IbpXiss_1 |   .9906573   .0161779    -0.57   0.565     .9594513    1.022878
         rts |   .4779762   .0289156   -12.20   0.000     .4245337    .5381464
  _IbpXrts_1 |   .7347408   .1127582    -2.01   0.045     .5438801    .9925791
  _IbpXage_1 |    .872296   .0932179    -1.28   0.201     .7074574    1.075542
      _Ibp_1 |   .0760752   .3033795    -0.65   0.518     .0000307    188.6858
_IbpXagesq_1 |   6.322103   8.321638     1.40   0.161     .4791204     83.4216
------------------------------------------------------------------------------
. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        11.61
                 Prob > chi2 =         0.1695
```

We will look shortly at what is being fit in terms of age, but note how much larger is the p-value for the goodness-of-fit test. We should be ready to proceed with reducing the model.

I will consider a backward elimination with $\alpha = .10$. Following the hierarchy principle, the only candidates for exclusion at step 1 are the interactions. Each of the 5 main effects is

involved in 1 or more interactions, so we cannot eliminate any main effects initially. The least significant interaction is BP*ISS with a p-value of .565, so this effect is removed (.565 > .10).

1. Omit BP*ISS and fit the model

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5 \sqrt{\text{Age}} + \beta_7(\text{BP} * \text{RTS}) + \beta_8(\text{BP} * \text{Age}) + \beta_9(\text{BP} * \sqrt{\text{Age}})$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*rts i.bp*age i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*age          _IbpXage_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(8)      =     944.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -440.39915                       Pseudo R2       =     0.5175
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070782    .007818     9.37   0.000     1.055568    1.086216
         age |   1.109481   .0254389     4.53   0.000     1.060726    1.160478
     agesqrt |   .4963908   .1314641    -2.64   0.008     .2953871    .8341726
         rts |   .4751856   .0283819   -12.46   0.000     .4226906       .5342
  _IbpXrts_1 |   .7451743     .11272    -1.94   0.052     .5539868    1.002343
  _IbpXage_1 |   .8700633   .0941304    -1.29   0.198     .7038191    1.075575
      _Ibp_1 |   .0480244   .1897495    -0.77   0.442     .0000208    110.8277
_IbpXagesq_1 |   6.604057   8.789526     1.42   0.156     .4863207    89.68069
------------------------------------------------------------------------------
```

At this step the candidates for exclusion are ISS, BP*Age, BP*$\sqrt{\text{Age}}$, and BP*RTS, of which BP*Age is least significant with a p-value of .198. This interaction is then omitted. Why is ISS a candidate for exclusion at this point?

2. Omit BP*Age and fit

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5 \sqrt{\text{Age}} + \beta_7(\text{BP} * \text{RTS}) + \beta_9(\text{BP} * \sqrt{\text{Age}})$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*rts  i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(7)      =     942.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -441.34771                       Pseudo R2       =     0.5164
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070154   .0078046     9.30   0.000     1.054966    1.085561
         rts |    .475377   .0283584   -12.47   0.000     .4229219    .5343383
         age |   1.100623   .0245862     4.29   0.000     1.053474    1.149881
      _Ibp_1 |   5.772836    6.82775     1.48   0.138     .5683829    58.63236
  _IbpXrts_1 |   .7505693   .1117116    -1.93   0.054     .5606625    1.004801
     agesqrt |   .5432324   .1411161    -2.35   0.019     .3264886    .9038644
_IbpXagesq_1 |   1.228323    .213722     1.18   0.237     .8733895    1.727497
------------------------------------------------------------------------------
```

At this step the candidates for exclusion are ISS, BP*$\sqrt{\text{Age}}$, and BP*RTS, of which BP*$\sqrt{\text{Age}}$ is least significant with a p-value of .237. This interaction is then omitted.

3. Omit BP*$\sqrt{\text{Age}}$ and fit

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5 \sqrt{\text{Age}} + \beta_7 (\text{BP} * \text{RTS})$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*rts
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*rts          _IbpXrts_#          (coded as above)
note: _Ibp_1 dropped due to collinearity
note: rts dropped due to collinearity
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(6)      =     941.24
                                                  Prob > chi2     =     0.0000
Log likelihood =  -442.0633                       Pseudo R2       =     0.5156
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070038    .007773     9.32   0.000     1.054912    1.085382
         rts |   .4705327   .0280666   -12.64   0.000     .4186169    .5288869
         age |   1.102008   .0246182     4.35   0.000     1.054798     1.15133
     agesqrt |    .550232   .1433209    -2.29   0.022     .3302406     .916772
      _Ibp_1 |   13.35317   12.52076     2.76   0.006     2.125423    83.89257
  _IbpXrts_1 |   .7906284   .1090461    -1.70   0.089     .6033535    1.036032
------------------------------------------------------------------------------
```

The candidates for exclusion at this point are ISS, Age, $\sqrt{\text{Age}}$, and BP*RTS. The least significant effect is BP*RTS with a p-value of .089, which is less than our criterion of .10.

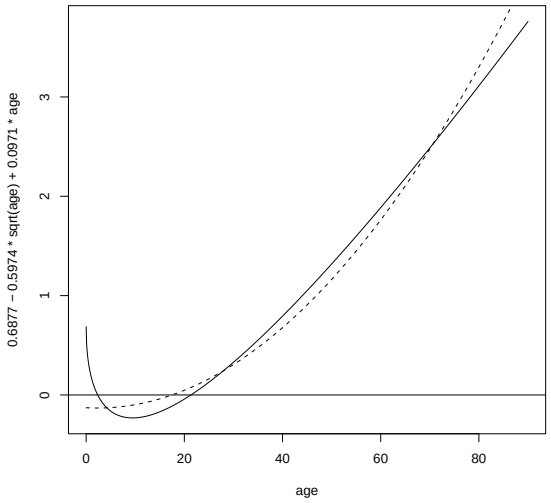
4. If we stick to the algorithm, we would stop and conclude that the important predictors are ISS, BP, RTS, AGE, and $\sqrt{\text{Age}}$ with an interaction between BP and RTS. All these steps can be automated with `sw`, as in the following output. I used logit here in order to see coefficients. Since I cannot combine `xi` and `sw`, I need to use `xi` alone to create indicator variables and then use those in `sw` for variable selection. `lockterm1` forces the first term in parentheses to stay in the model.

```
. xi i.bp i.bp*iss i.bp*rts i.bp*age i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*iss          _IbpXiss_#          (coded as above)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*age          _IbpXage_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)

. sw  logit death (iss  _Ibp_1 rts  age agesqrt)   _IbpXiss_1  _IbpXrts_1 _IbpXage_1
>  _IbpXagesq_1, pr(.1) lockterm1
                    begin with full model
p = 0.5654 >= 0.1000  removing _IbpXiss_1
p = 0.1983 >= 0.1000  removing _IbpXage_1
p = 0.2372 >= 0.1000  removing _IbpXagesq_1
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(6)      =     941.24
                                                  Prob > chi2     =     0.0000
Log likelihood =  -442.0633                       Pseudo R2       =     0.5156
------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   .0676945   .0072642     9.32   0.000      .053457    .0819321
      _Ibp_1 |   2.591754   .9376617     2.76   0.006     .7539709    4.429537
         rts |  -.7538899   .0596486   -12.64   0.000    -.8707991   -.6369807
         age |   .0971337   .0223394     4.35   0.000     .0533494    .1409181
     agesqrt |  -.5974152   .2604735    -2.29   0.022    -1.107934   -.0868965
  _IbpXrts_1 |  -.2349272   .1379234    -1.70   0.089     -.505252    .0353977
       _cons |   .6877421   .8100227     0.85   0.396    -.8998732    2.275357
------------------------------------------------------------------------------

. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        14.46
                 Prob > chi2 =         0.0704
```

The fitted model is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = .688 + .068 \text{ ISS} + 2.59 \text{ BP} - .754 \text{ RTS} + .097 \text{ Age} - .597\sqrt{\text{Age}} - .235 \text{ BP} * \text{RTS}$$

The regression effect for ISS is easily interpreted as a risk factor for death (why?). The effect of age needs to be examined graphically since it is not simply linear. In the plot below the solid line is for the fitted model above, and the dotted line is what happens if we use AGE and AGE^2 instead. Can you see why I preferred using $\sqrt{\text{AGE}}$ to AGE^2? The fitted model shows increased risk of death for very young children, lowest risk for children and young adults, and substantially increased risk for older adults.



The effects of BP and RTS are more difficult to interpret because they interact. For example, for any fixed ISS and Age,

$$\begin{aligned}
\widehat{OR} &= \frac{\widehat{\text{odds}} \text{ of death for BP=1 (Penetrating)}}{\widehat{\text{odds}} \text{ of death for BP=0 (Blunt)}} \\
&= \frac{e^{.688+.068\text{ISS}+2.59(1)-.754\text{RTS}+.097\text{Age}-.597\sqrt{\text{Age}}-.235(1)\text{RTS}}}{e^{.688+.068\text{ISS}+2.59(0)-.754\text{RTS}+.097\text{Age}-.597\sqrt{\text{Age}}-.235(0)\text{RTS}}} \\
&= e^{2.59-.235\text{RTS}}
\end{aligned}$$

which decreases for increasing RTS. Looking at the ends of the RTS spectrum,

	RTS	$\widehat{OR}$
(no vitals)	0	13.35
(normal)	7.84	2.12

So, depending on ones RTS, the estimated odds of dying from a penetrating injury vary from 2 to 13 times the odds of dying from a blunt trauma, adjusting for ISS and Age. Before jumping on this large difference very hard, though, let's look at confidence intervals, which do overlap quite a bit here.

```
. lincom _b[_Ibp_1],or
 ( 1)  _Ibp_1 = 0
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   13.35317   12.52076     2.76   0.006     2.125423    83.89257
```

```
------------------------------------------------------------------------------
. lincom _b[_Ibp_1]+7.84*_b[_IbpXrts_1],or
 ( 1)  _Ibp_1 + 7.84 _IbpXrts_1 = 0
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.116841    .628437     2.53   0.012     1.183009    3.787814
------------------------------------------------------------------------------
```

Remarks

1. Some epidemiologists force *confounders* to be included in a logistic regression model regardless of their statistical significance.

2. The BP*RTS interaction was barely significant at the $\alpha = .10$ level. It might be interesting to see whether ones conclusions change when this effect is omitted.

```
. xi:logistic death iss i.bp rts age agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(5)      =     937.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -443.73652                       Pseudo R2       =     0.5138
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.069022   .0077489     9.21   0.000     1.053942    1.084318
      _Ibp_1 |   2.911404   .6711942     4.64   0.000     1.852963    4.574443
         rts |   .4474509   .0239457   -15.03   0.000     .4028958    .4969333
         age |   1.102117   .0249277     4.30   0.000     1.054327    1.152073
     agesqrt |   .5548027   .1463495    -2.23   0.026     .3308286    .9304094
------------------------------------------------------------------------------
. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        14.77
                 Prob > chi2 =         0.0637
```

We see that remaining effects are highly significant and there is no evidence of gross deficiencies.

3. The $\widehat{OR}$s for ISS and Age are similar for the two models. If a primary interest was estimating OR for ISS or Age, then it would not matter much which model we used. If BP is the interesting effect, the simpler model yields an $\widehat{OR}$ of 2.91, which is between the minimum and maximum $\widehat{OR}$ for the previous model.

4. The model without BP*RTS is simpler to interpret because it contains no interactions. However, most scientists are wary of omitting potentially important interactions, because of the potentially misleading conclusions that might be reached in models that ignore them. I would be inclined here to use the slightly more complex model with the BP*RTS interaction.

Case-Control Data

In epidemiological studies, the logistic model $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is used to relate p, say the probability of disease or death, to the collection $x_1, x_2, \ldots, x_k$ of risk factors and confounders. With prospective or cross-sectional studies, we have noted that risk (i.e. probability of disease or death), relative risks, and ORs can be estimated using the logistic model – however most of our focus has been on ORs.

In practice, data are often sampled retrospectively using a case-control design. Although it is well known that risks and relative risks cannot be estimated using case-control data, ORs are

estimable and agree with ORs defined from a prospective study. In terms of logistic regression, the intercept cannot be estimated without bias using data from a case-control study, but regression coefficients for predictors and confounders, which correspond to adjusted ORs, are estimated appropriately. Thus, we can use standard methods to estimate regression effects and build regression models using case-control data.

Diverticular Disease Example
There is a description of this data set on the web page as a supplement to this lecture. The data set also is provided there. The data set has 64 rows with this content:

Variable Name	Contents
Age	Midpoint of age range (8 levels)
Sex	Values are f and m
Cancer	Colonic Cancer (1 is yes - case, 0 is no - control)
Lab	Case - Control label (not used)
Disease	Diverticular disease (values dd (yes) and ndd (no))
Count	No. individuals with this combination of variables

There are a lot of possible strategies for building a model to predict Cancer. I proceeded this way:

1. The primary interest is the potential association with diverticular disease (DD) and colonic cancer (CC). DD is considered an exposure variable.

2. Age and sex are viewed as confounders (potentially). Confounders are variables that are risk factors for the disease and associated with, but not a consequence of, presence or absence of the exposure variable.

   Because age and sex are likely to confound the relationship between the occurrence of DD and CC, most epidemiologists would argue that the effect of DD has to be assessed after adjusting for the effects of age and sex. As a result, many epidemiologists would include age and sex effects in a model, regardless of their statistical significance. Others might adopt a slightly different view and consider the effect of removing insignificant sex and age effect on adjusted ORs for DD. If removing insignificant effects has little impact on the estimate and precision of the adjusted OR for DD it does not matter much whether they are included or excluded. If the adjusted OR for DD changes dramatically upon removal, the insignificant effect would typically remain in the model.

3. We have the option of treating Age, using midpoint of the age range, as a categorical variable or on a continuous scale. If we consider Age as categorical, the odds of CC will be allowed to vary freely across age categories – that is, the odds is not required to vary smoothly with Age. If we choose this approach, interpretation of Age effects and interactions with Age will be cumbersome. However, almost every logistic model with Age, Sex, and DD effects fits well (using goodness of fit measures) when Age is categorical but fits poorly when Age is continuous. This implies that the log odds of CC does not change linearly with Age, but follows a more complex pattern. Consequently, I considered adding a quadratic term in Age, and this improved the fit dramatically.

4. I then posed a full model with the following effects: Sex, DD, Age, Age^2, Sex*Age, Sex*DD, Age*DD. I then proceeded with a Backward Elimination. I decided to force DD, Sex, and Age to be included in the model, regardless of their significance, but all other effects were candidates for exclusion. Note: Count must be defined as a frequency variable.

5. **Stata** is not going to let us use character variables directly in `logistic`, but that's no problem here since we need to create appropriate indicator variables and interactions anyway. `xi` is accommodating, though, so first we generate the indicators and then perform the sw procedure with the constraints listed above.

```
 . xi i.sex i.disease i.sex*age i.sex*i.disease i.disease*age
 i.sex             _Isex_1-2           (_Isex_1 for sex==f omitted)
 i.disease         _Idisease_1-2       (_Idisease_1 for disease==dd omitted)
 i.sex*age         _IsexXage_#         (coded as above)
 i.sex*i.disease   _IsexXdis_#_#       (coded as above)
 i.disease*age     _IdisXage_#         (coded as above)
. sw logit  cancer (age _Isex_2 _Idisease_2) agesq _IsexXage_2  _IsexXdis_2_2
        _IdisXage_2 [fweight=count],pr(.1) lockterm1
                      begin with full model
p = 0.4841 >= 0.1000  removing _IsexXdis_2_2
Logit estimates                                   Number of obs   =        193
                                                  LR chi2(6)      =      46.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -104.72637                       Pseudo R2       =     0.1818
------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.6267873   .2185572    -2.87   0.004    -1.055151   -.1984232
     _Isex_2 |   4.009058   2.022184     1.98   0.047     .0456503    7.972465
 _Idisease_2 |  -4.604635   3.183622    -1.45   0.148    -10.84442    1.635149
       agesq |   .0053892   .0016092     3.35   0.001     .0022352    .0085432
 _IsexXage_2 |  -.0737373   .0318366    -2.32   0.021    -.1361359   -.0113386
 _IdisXage_2 |   .0806418   .0479469     1.68   0.093    -.0133323    .1746159
       _cons |   16.93369   7.510259     2.25   0.024     2.213853    31.65353
------------------------------------------------------------------------------
```

The `lockterm1` option forces `(age _Isex_2 _Idisease_2)` to stay in the model no matter what. Only the DD*Sex interaction term was removed in the backward elimination, so we have a model left with Age, Sex, DD, Age^2, Age*Sex, and Age*DD effects.

6. The goodness of fit test for the final model shows no problems.

```
Logistic model for cancer, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =       193
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        10.80
                Prob > chi2 =         0.2131
```

7. The parameter estimates table is given only for the final model when using `sw`.

8. A primary interest is the effect of disease on CC. `xi` produced `_Idisease_2` and `_IdisXage_2` where `_Idisease_2` is 1 for ndd, 0 for dd; and `_IdisXage_2` is 0 for dd and Age for ndd. We want to measure odds of cancer for ndd and dd. Using the same reasoning as previously (write the model, cancel common terms – the ones that are the same for dd and ndd),

$$\widehat{OR}\text{ (NDD vs. DD)} = e^{-4.604635+.0806418\text{Age}}$$

We could use this formula directly, but it is considerably easier to use `lincom` as before. I just computed the estimated OR for each of the ages in the data set, with the following results.

```
. lincom _b[ _Idisease_2]+44.5*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 44.5 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3620128   .3995044    -0.92   0.357     .0416264    3.148324
------------------------------------------------------------------------------
. lincom _b[ _Idisease_2]+52*_b[  _IdisXage_2],or
```

```
 ( 1)  _Idisease_2 + 52 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .6628131   .5182929    -0.53   0.599     .1431482    3.068995
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+57*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 57 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .991979   .5875997    -0.01   0.989     .3106651     3.16747
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+62*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 62 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.484615   .6725879     0.87   0.383     .6109234    3.607788
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+67*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 67 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.221904    .928302     1.91   0.056     .9797086    5.039108
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+72*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 72 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.325345   1.691708     2.36   0.018     1.226884    9.013008
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+77*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 77 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   4.976776   3.368095     2.37   0.018     1.320952    18.75035
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+84.5*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 84.5 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   9.112032   8.985077     2.24   0.025     1.319086    62.94442
------------------------------------------------------------------------------
```

The confidence intervals indicate that OR really doesn't differ significantly between DD and NDD for patients under 70, but for older patients DD appears actually to be protective. We should check to see if this is a real pattern in the data, or a fluke of the model we have fit. How would you do such an analysis? We also need to make sure it makes some sense to someone who knows the medicine.

I hope you see the value in including terms like Disease in the model, even though it is not actually significant in this case. We needed to assess the potential for this variable to affect CC through adjusted ORs, and we did find an interesting relationship (because age is so important).

As an aside, I will note that if you remove the effect for Sex (and its interaction with age), this has little effect on adjusted OR for DD. If age is completely ignored in the analysis the adjusted OR for DD is reduced dramatically, implying that age is clearly an important confounding variable in the relationship between DD and CC.

You can calculate any estimated adjusted OR using the above method. Remember, however, that this is a case-control study, so risks or odds should not be evaluated!

In-Lab Exercise

Return to the Framingham study data. Run the following code (make sure you understand what I am doing here):

```
graph bar chd [fw=freq],over(scl, ///
               relabel(1 "<190" 2 "190-219" 3 "220-249" 4 "250+")) ///
               over(agegroup, relabel(1 "30-49" 2 "50-62")) ///
               over(gender, relabel(1 "Female" 2 "Male")) ///
               ytitle("Proportion CHD") ///
               title("CHD vs. Gender, Age, and SCL")
bysort gender agegroup:tabulate chd scl [fw=frequency],chi2 exp col
```

Examine the output of the bar graphs and chi-squared tests.

1. What main effects appear to be present?
2. What interactions appear to be present?
3. Find a suitable model using logistic regression.
4. Summarize important odds ratios from your logistic regression model.
5. Give an overall summary of the analysis.