Models In Epidemiology And Biostatistics
Gordon Hilton Fick

A Detailed Example

Lets take a look now at a cohort study designed to assess the relationship between a baseline catecholamine level classified as low (cat=0) or high (cat=1) and the development of coronary heart disease (chd=1) over a 9 year follow up period. This study included a group of 609 white males. Some baseline characteristics were measured: age in years (age), cholesterol level in US units of mg/dL (chl), systolic blood pressure in mm Hg (sbp), diastolic blood pressure in mm Hg (dbp) while some baseline characteristics were dichotomous: electrocardiogram abnormal status (ecg=1 means abnormal), hypertension status (hpt=1 means hypertensive) and smoking status (smk=1 means smoker). The data is in evans.dta

Accordingly we will consider that disease status is chd and exposure status is cat and that all the other variables are viewed as potential confounders and/or modifiers. Further, it is suspected that there may be elaborate forms of modification or confounding. You might note that it is usually the preference to study a rate ratio with this sort of study but we will pursue an analysis via odds ratios. We will return to this matter in a later class.

Maybe we should switch the units of [total] cholesterol to metric (mmol/L) by dividing the US values by 38.6 but it turns out that such a linear transformation has no effect on the p-values. When we are interpreting "per unit change" statements, then there will apparent differences since a unit in metric is 38.6 units in US. Such differences are artifactual though. It has been reported that chl [in US units] >200 is "bad" which converts roughly to chl [in metric] > 5 is "bad". Maybe a cutoff (or cutoffs) should be considered.

Ignoring the measured characteristics still gives us 8 2x2 tables:

```
. bysort smk ecg hpt:tab chd cat

-> smk = no, ecg = nor, hpt = nor

           |    cat
       chd |       low |     Total
-----------+-----------+----------
    no chd |       103 |       103
       chd |         2 |         2
-----------+-----------+----------
     Total |       105 |       105

-> smk = no, ecg = nor, hpt = hyp

           |          cat
       chd |       low       high |     Total
-----------+----------------------+----------
    no chd |        40          9 |        49
       chd |         3          1 |         4
-----------+----------------------+----------
     Total |        43         10 |        53

-> smk = no, ecg = abn, hpt = nor

           |          cat
       chd |       low       high |     Total
-----------+----------------------+----------
    no chd |        14          5 |        19
       chd |         3          1 |         4
```

```
-----------+----------------------+----------
     Total |        17          6 |        23

-> smk = no, ecg = abn, hpt = hyp

           |          cat
       chd |       low       high |     Total
-----------+----------------------+----------
    no chd |        11         23 |        34
       chd |         1          6 |         7
-----------+----------------------+----------
     Total |        12         29 |        41

-> smk = yes, ecg = nor, hpt = nor

           |          cat
       chd |       low       high |     Total
-----------+----------------------+----------
    no chd |       164          6 |       170
       chd |        11          4 |        15
-----------+----------------------+----------
     Total |       175         10 |       185

-> smk = yes, ecg = nor, hpt = hyp

           |          cat
       chd |       low       high |     Total
-----------+----------------------+----------
    no chd |        57         22 |        79
       chd |        16          5 |        21
-----------+----------------------+----------
     Total |        73         27 |       100

-> smk = yes, ecg = abn, hpt = nor

           |          cat
       chd |       low       high |     Total
-----------+----------------------+----------
    no chd |        29          5 |        34
       chd |         4          3 |         7
-----------+----------------------+----------
     Total |        33          8 |        41

-> smk = yes, ecg = abn, hpt = hyp

           |          cat
       chd |       low       high |     Total
-----------+----------------------+----------
    no chd |        25         25 |        50
       chd |         4          7 |        11
-----------+----------------------+----------
     Total |        29         32 |        61
```

One can see that cell numbers are small in some strata [even without "accounting" for chl or age] and, in particular, no one had a high cat in the strata of normotensive nonsmokers with normal ecg. Lets try a [very provisional] stratified analysis.

```
. egen hse=group(hpt smk ecg)
. lab def hsel 1 "nor_no_nor" 2 "nor_no_abn" 3 "nor_yes_nor" 4 "nor_yes_abn" 5 "hyp_no_nor"
6 "hyp_no_abn" 7 "hyp_yes_nor" 8 "hyp_yes_abn"

. lab val hse hsel
```

```
. cc chd cat,by(hse)

group(hpt smk ec |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
      nor_no_nor |          .              .          .            0 (exact)
      nor_no_abn |   .9333333      .0148468   15.32852     .6521739 (exact)
     nor_yes_nor |   9.939394      1.748926   48.45572     .3567568 (exact)
     nor_yes_abn |       4.35      .4674848   34.43807     .4878049 (exact)
      hyp_no_nor |   1.481481      .0254274    20.9407      .509434 (exact)
      hyp_no_abn |   2.869565      .2828535   143.7862     .5609756 (exact)
     hyp_yes_nor |   .8096591      .2070997    2.69632         3.52 (exact)
     hyp_yes_abn |       1.75      .3828657   9.133987     1.639344 (exact)
-----------------+-------------------------------------------------
           Crude |   2.861483      1.614858   4.987845              (exact)
    M-H combined |   1.858531      1.030747   3.351101
-----------------------------------------------------------------------
Test of homogeneity (B-D)      chi2(7) =    10.10  Pr>chi2 = 0.1830

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      4.53
                                                Pr>chi2 =    0.0332
```

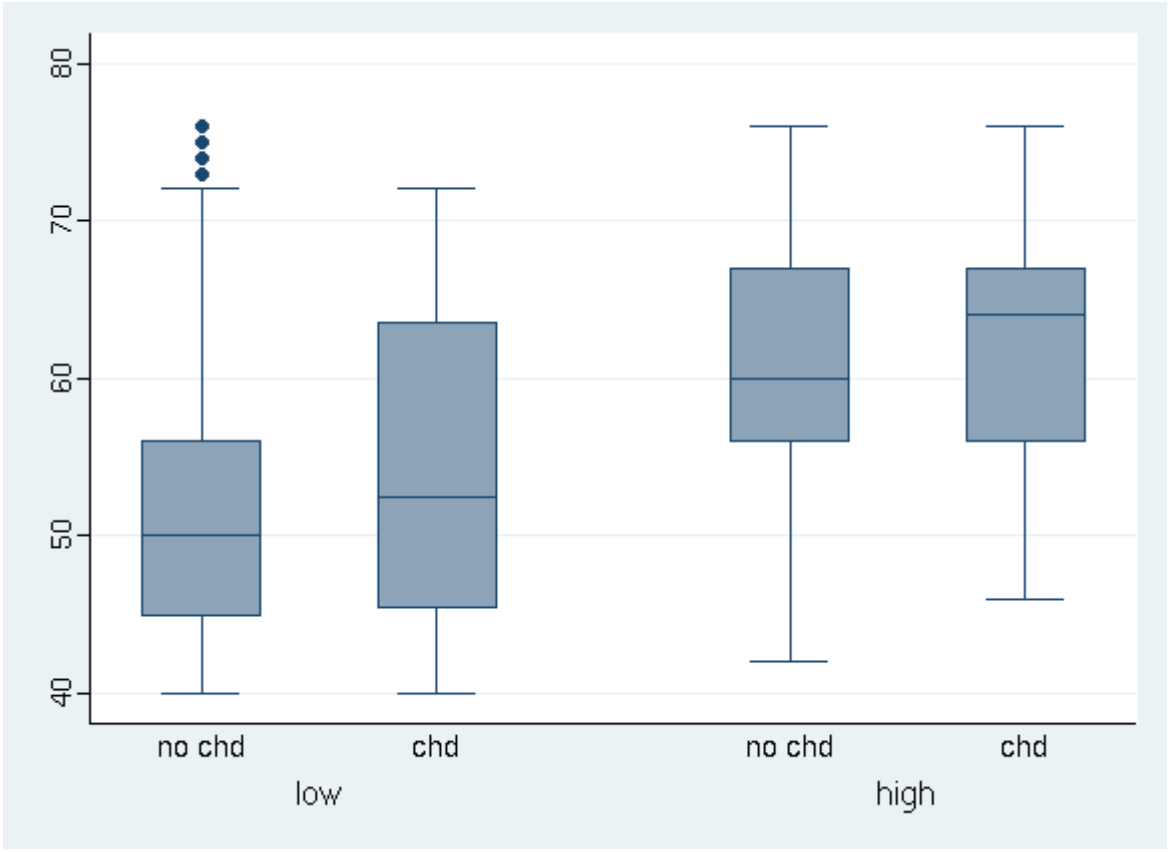
One might glean from this "analysis" that hpt and smk may modify [why?] but this is pretty rough. Maybe we should toss out ecg for the time being.

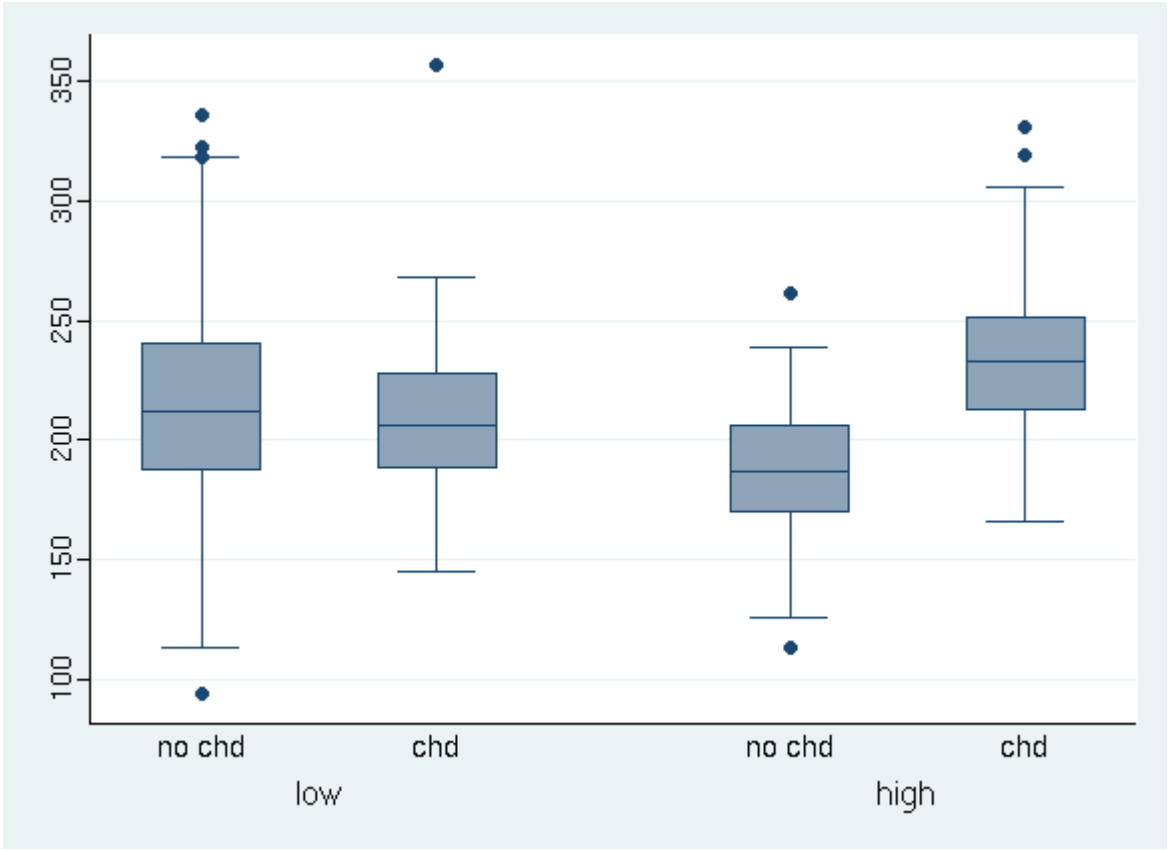
```
. egen hs=group(hpt smk)

. lab def hsl 1 "nor_no" 2 "nor_yes" 3 "hyp_no" 4 "hyp_yes"

. lab val hs hsl

. cc chd cat,by(hs)

  group(hpt smk) |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
          nor_no |       4.68      .0826078   55.28377     .1953125 (exact)
         nor_yes |   8.187879      2.299428   26.96322     .7300885 (exact)
          hyp_no |   2.789063      .6409602   13.90651     1.361702 (exact)
         hyp_yes |   1.046809      .4260243   2.486613     5.838509 (exact)
-----------------+-------------------------------------------------
           Crude |   2.861483      1.614858   4.987845              (exact)
    M-H combined |   2.067735      1.185968   3.605096
-----------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(3) =     9.65  Pr>chi2 = 0.0218

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      7.46
                                                Pr>chi2 =    0.0063
```

Looks like hpt and smk modify but we have not even considered the measured variables yet: age,chl, sbp and dbp.
Maybe hpt is covering the potential value of the measured blood pressure values, particularly if the hpt classification is based on the measured values and any anti-hypertensive meds a participant is taking. Anyhow, we will toss sbp and dbp here.

Graphs of chl versus age (for each cat/chd combination) reveal no patterns or apparent association. So multicollinearity may not be an issue. [More this matter on future classes]
Lets look at age and chl ignoring hpt, smk and ecg with some graphical assessment.

```
graph box age, over(chd) over(cat)
graph box chl, over(chd) over(cat)
```


80
70
60
50
40
no chd
chd
no chd
chd
low
high


350
300
250
200
150
100
no chd
chd
no chd
chd
low
high

Higher ages for those with chd compared to those without chd. Big surprise. Looks like higher chl for those with chd compared to those without chd but, maybe, only for those with high cat.

Looks like both age and chl are involved in the chd/cat relationship but this rough assessment has ignored smk, hpt and ecg.

So.... the data is entered and cleaned. A huge effort, typically. Maybe months or even years have been spent getting this far. Tables and graphs have been considered. The literature has been picked over for any clues as to what should happen next in the analysis. We now enter the stage where serious health research can involve lengthy assessments of various candidate models through interpreting the models, debating the merits and demerits. Trying to simplify a model or realizing that what you have is far too simple.

There are some basic rules [maybe basic guidelines is a better phrase?] out there. In some content areas, certain data transformations are well established for certain variables. [for example: duration] Also, it may be agreed that thresholds must at least be considered. [for example: macroalbuminuria]. Such thresholds are perhaps so ingrained in a content area, that colleagues expect to see their consideration, at least.

Now what? A model for the log of the odds of chd with cat, smk, hpt, ecg, along with a linear component for age and chl has 6 factors and if we included all interactions would have $2^6 = 64$ terms! If we tried to fit this model, Stata would object (and so would I!). There are "automated" procedures for model selection out there in the pseudo-statistics literature. You may know of "stepwise" methods. [they are available in Stata] There are serious criticisms of such stepwise procedures and it is generally agreed that so-called "expert systems" must be viewed with cynicism and careful review. It is worth noting that so-called "data mining" has achieved some respectability these days [the whole field of micro-arrays being the most noted example] although hard-core scientists remain very skeptical.

So... back to the issue at hand. How do we proceed then? We have to make some choices. A simple additive model (with no interactions) is out of the question. A model with 64 terms is also out. What then?

How about some middle ground? Whatever that means. Well, following a generally accepted stratified analysis method paradigm, we should begin with the complex and try to move to the simpler (sometimes unpleasantly called "backward elimination") rather than starting with the simple and trying to make the model more complex (so called "forward selection") although both strategies have their merits. I will try to illustrate this challenge by using a bit of both.

Maybe a decent place to start is with more than one model with components somehow compartmentalized (is that a real word?) like:

A logistic model to reproduce the stratified analysis in C=cat, S=smk and H=hpt

$$\log(p/(1-p)) = \beta_0 + \beta_1 C + \beta_2 S + \beta_3 H + \beta_4 CS + \beta_5 CH + \beta_6 SH + \beta_7 CSH$$

along with 2 simple models: one with A=age and one with L=chl

$$\log(p/(1-p)) = \beta_0 + \beta_1 C + \beta_2 A + \beta_3 CA$$

$$\log(p/(1-p)) = \beta_0 + \beta_1 C + \beta_2 L + \beta_3 CL$$

Serious learning can come from these models and an acclimatization to the issues, the nature of the relationships and the possible complexities that may make sense. We not proposing that considering these three models as being anywhere near our goal. But maybe such construction and assessment is at least a start.

It cannot be the intention here to provide a lengthy comprehensive set of analyses, but lets get the ball rolling by considering a slightly more elaborate model and begin the assessment process. For no very good reason, lets ignore smk and chl and try:

```
. gen agec=age-53.7
. gen ha=agec*hpt
. gen ca=agec*cat
. gen ch=cat*hpt
. gen cha=ca*hpt

. logit chd cat hpt agec ch ca ha cha

Logistic regression                                     Number of obs   =        609
                                                        LR chi2(7)      =      37.11
                                                        Prob > chi2     =     0.0000
Log likelihood = -200.72472                             Pseudo R2       =     0.0846

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cat |   1.140279   .7904288     1.44   0.149    -.4089333    2.689491
         hpt |   1.034008    .331452     3.12   0.002      .384374    1.683642
        agec |  -.0045154    .027762    -0.16   0.871     -.058928    .0498971
          ch |  -.9045976   .9020882    -1.00   0.316    -2.672658    .8634627
          ca |   .1523009   .0836721     1.82   0.069    -.0116933    .3162952
          ha |   .0603245   .0362756     1.66   0.096    -.0107743    .1314234
         cha |  -.2012501   .0926563    -2.17   0.030    -.3828531   -.0196471
       _cons |  -2.750515   .2392519   -11.50   0.000     -3.21944    -2.28159
------------------------------------------------------------------------------

. predict lp, xb
. sort cat hpt age
. twoway (line lp age if cat==0 & hpt==0) (line lp age if cat==1 & hpt==0) (line lp age if
cat==0 & hpt==1) (line lp age if cat==1 & hpt==1),legend(order (1 "cat==0 & hpt==0" 2
"cat==1 & hpt==0" 3 "cat==0 & hpt==1" 4 "cat==1 & hpt==1"))
```

The model here is:

$$\log(p/(1-p)) = \beta_0 + \beta_1 C + \beta_2 H + \beta_3 A + \beta_4 CH + \beta_5 CA + \beta_6 HA + \beta_7 CHA$$

A graph [below] of the fits is four lines.

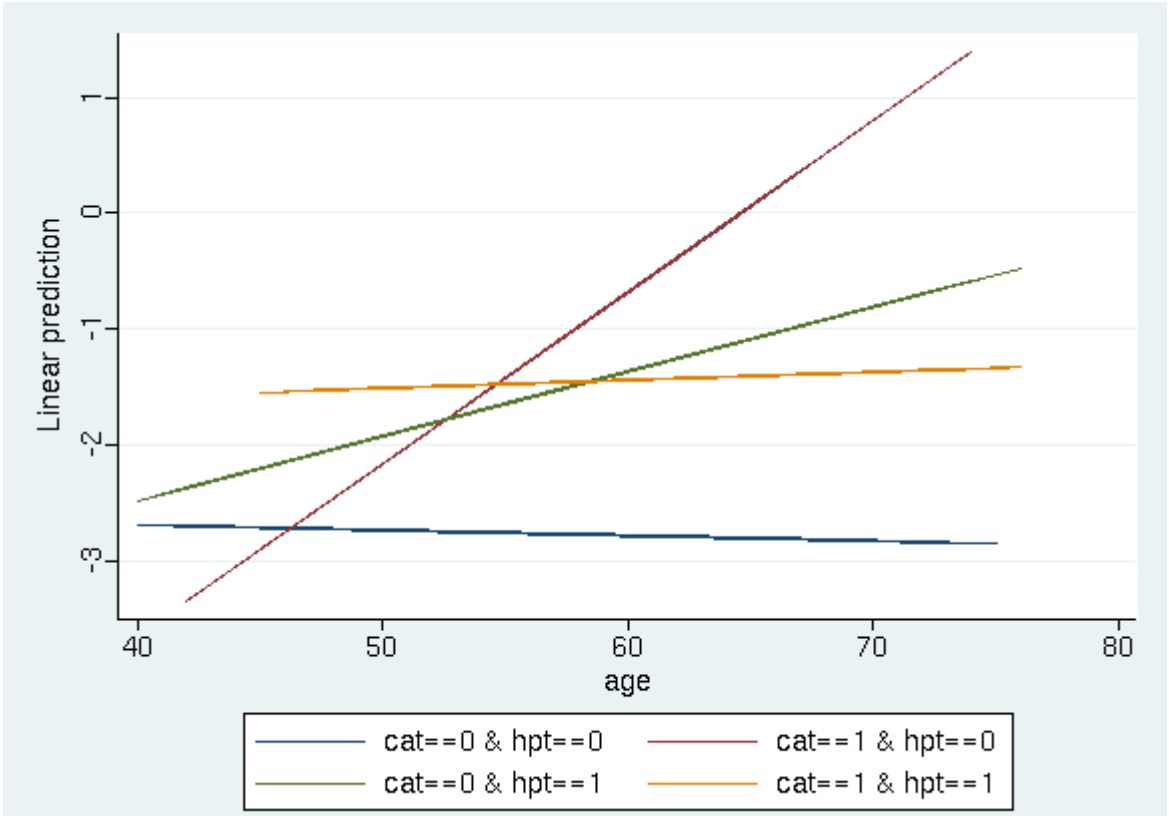
From this model, for normotensives, we get: $\log(\text{OR}) = \beta_1 + \beta_5 A$

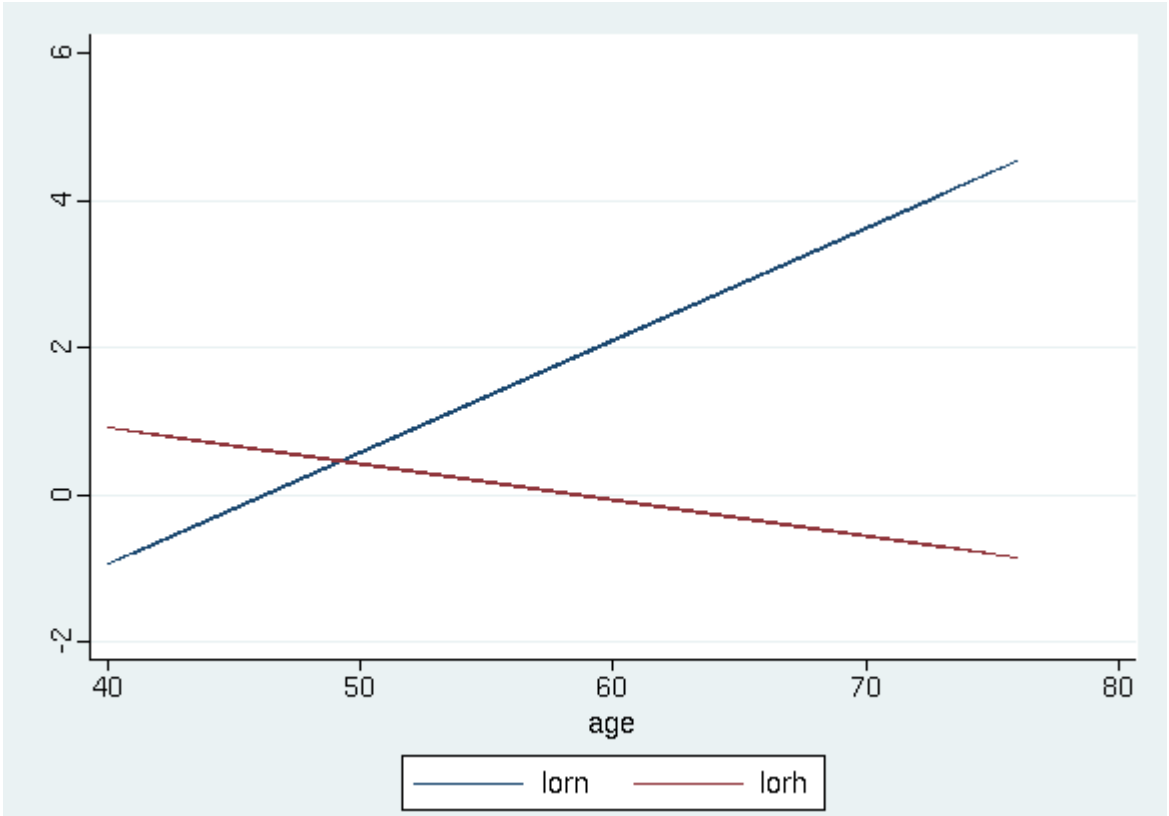
while for the hypertensives, we get: $\log(\text{OR}) = \beta_1 + \beta_4 + (\beta_5 + \beta_7) A$

so that β_7 provides a measure of how age modification depends on hypertensive status. Since the p-value associated with this coefficient is 0.030, it would appear that age modification does depends on hypertensive status.

A graph of these two lines is included after the 4 line plot using:

```
. gen lorn =1.140279 + 0.1523007*agec
. gen lorh =(1.140279-0.9045976) + (0.1523007-0.2012501)*agec
. line lorn lorh age
```


Linear prediction
1
0
-1
-2
-3
40
50
60
70
80
age
cat==0 & hpt==0
cat==1 & hpt==0
cat==0 & hpt==1
cat==1 & hpt==1


6
4
2
0
-2
40
50
60
70
80
age
lorn
lorh

It would appear that the log(OR) rises with age for the normotensives but the log(OR) may not depend on age for the hypertensives. Maybe, the next step would be to include smk and/or chl in the model to see if this relationship is maintained.

The above pages are only the first steps in a careful analysis.