Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 10 : Issues When Modeling

This session provides an introduction to a number of different issues. These issues can appear quite often when using models.

Linearly Dependent Columns

All of regression is about linear combinations (or weighted sums) of columns of data: $\sum \beta_i x_i$ The weights are the $\beta_i's$ while the columns are the $x_i's$ You may recall that such columns can be thought of as vectors. [column vectors!] All of the basics from linear algebra may come rushing back to you just now. [maybe not...]. The $\sum \beta_i x_i$ provides us with a 'log of odds' in the context of logistic regression. It turns out that no subset of these vectors can form what is called a linearly dependent set.

To motivate linear dependency, let us suppose that we have a potential confounder/modifier that is characterized by levels. For example, age group with 3 'levels': young, middle aged and old with codes 1, 2 and 3 respectively. Think now of the indicator variable for each of these levels: a_1, a_2, a_3 where a_i is 1 if the participant is in group i and is 0 otherwise. Now notice that $a_1+a_2+a_3=1$ Check out what this means for a particular participant. If $a_1=0$ and $a_2=0$, then we know that a_3 must be one. Meaning: if we know a participant is not young and not middle aged, then, we know they are old. Here, we are assuming that the characteristic 'age group' is made up of a mutually exclusive and exhaustive set of levels. Each participant is described by 'age group' in one and only one of these levels. Knowing any 2 of the a_i values determines the third. [For sure]. We then have four columns [vectors] that are said to be linearly dependent.

The actual definition is a little technical:

A set of p+1 vectors $a_0, a_1, a_2 \ldots, a_p$ form a linearly dependent set if there are $c_i's$ [scalars not all equal to zero] so that $\sum_{i=0}^{p} c_i a_i = 0$. If there are no such $c_i's$, then we say that these vectors are linearly independent.

For our example above, p=3, $a_0=1$ and so we can pick $c_0=-1,\ c_1=1,\ c_2=1$ and $c_3=1$ which then gives us $a_1+a_2+a_3=1$. Since there are non-zero $c_i's$, we know that $1,\ a_1,\ a_2$ and a_3 are linearly dependent. We will see that if we exclude any one of $1,\ a_1,\ a_2$ and a_3 , then the remaining three form a linearly independent set. Indeed, such a step makes sense once we interpret the coefficients that result.

Lets consider, as a starting place: $\log(p/(1-p))=\beta_0+\beta_1 E+\beta_2 a_2+\beta_3 a_3$ As always, one should interpret the coefficients. [In particular, interpret β_1]

Now, if we were to attempt to fit:
$\log(p/(1-p))=\beta_0+\beta_1 E+\beta_2 a_2+\beta_3 a_3+\beta_4 a_1$, all software will exclaim 'warning' and delete one of the a_i from the equation list. Usually the last in the equation is deleted. In this illustration, a_1 would be removed, but, in fact, any one of the three could have been selected. This type of variable

deletion does not necessarily mean that a_1 is not needed in such a model construct but, rather, that you, as the thinking part of all of this must now think: “I have a problem with my logic” “I need to trace through all the steps that have led me to this model”. In this case, we can see that knowledge of any 2 of the levels of age group determines the third and so the 'estimation' process cannot be managed without a change on your part. Usually, you want to make a choice. In this example, such a choice will determine which level of age group becomes the 'baseline' level and provides an interpretation to our $\beta_i's$ as differences relative to the chosen baseline level.

Lets make this example a little more elaborate. Consider:
[Model 1]

$$\log(p/(1-p))=\beta_0+\beta_1 E+\beta_2 a_2+\beta_3 a_3+\beta_4 E a_2+\beta_5 E a_3$$

Write out this equation for the young, the middle aged and the old. Interpret the coefficients.

Now consider:
[Model 2]

$$\log(p/(1-p))=\beta_1 a_1+\beta_2 a_2+\beta_3 a_3+\beta_4 E a_1+\beta_5 E a_2+\beta_6 E a_3 \quad .$$

Write out this equation for the young, the middle aged and the old. Interpret the coefficients.

In Stata, to try Model 2, you include the explanatory variables as always but now you add the option, noconstant. For example:

```
logit dis a1 a2 a3 Ea1 Ea2 Ea3, noconstant
```

After the above command, you could still consider a comparison with the 'lincon' command. For example, to estimate the log odds ratio for the middle-aged minus the log odds ratio for the young you would enter:

```
lincon Ea2-Ea1
```

Both models 1 and 2 accomplish exactly the same task. The fitted values from Models 1 and 2 are identical. The regression coefficients in Model 2 can be used to compute the regression coefficients in Model 1 by identifying their corresponding interpretations and vice versa.

In these situations, you always have choices. In our example, we could choose either one of the young, the middle-aged or the old as baseline. For a fourth option, we could choose not to have a baseline group and the construct the models using the noconstant option. It is always a good idea to check that you are making the right moves by constructing the fitted values for each alternative model and checking that the fitted values from each of the alternatives are the same.

“Nearly” Linearly Dependent Columns

The above example refers to exact linear dependent sets. In health research, we can have a set that is “close to being” a linearly dependent set. Sometimes, software will spit out one [or more] of the members from such a set [for removal] while in other circumstances, we may receive a fit with no deletions and not get a clear clue to trouble brewing. An example, may help to display the issue here.

Lets consider a study of diabetics in which the outcome is retinopathy [an eye disease that can lead to

blindness]. Lets suppose that we have recorded a patient's age A[in years], how long they have had diabetes D [in years] and their type of diabetes T [coding Type I =0 and Type II =1] For illustration only, lets us suppose we construct a model for the log of the odds of retinopathy using, say:

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

Notice that if T=0, then A and D are typically nearly the same. Type I diabetics are diagnosed at a young age so that here, D may be just a little less than A. While if T=1, we can see that A and D are typically different. In fact, if T=1, A and D are rarely close.

Notice that knowledge of T and A for a given participant, tells us 'a lot about' D. Certainly not exactly what D is but maybe 'a narrowing down' of the possible values for D.

This sort of phenomena is now usually called the 'multicollinearity' of A, D and T, in that knowledge of any 2 of A, D or T at least partially determines the third. [I have tried a search for history of the term 'multicollinearity'. We say points on a graph are collinear if they lie on a line. The shortest distance between 2 points is ....] [The term may go back to the mid 1960's when a researcher noticed that his data, using [an early algorithm of] SPSS, would give very different results with the same data run on different computers! The term 'ill-conditioned' was used then... maybe still is]

In practice, how do we avoid or detect multicollinearity?
Your content area literature may have such matters identified.
Even though a fit has been determined, some of the standard errors of the estimates may be far larger than one expects.
A coefficient may be in the wrong direction. For example, a negative value for the estimate of the coefficient for age may be a strong clue of trouble since we may know that the log odds of disease cannot decline with age.
Attempting to interpret a coefficient, may lead to an unrealistic scenario. In our example, if we attempt to interpret the coefficient of duration, we are 'fixing' type and age and then conceptualizing a rate of change of log of odds of disease per year of duration. This may be fantasy. For when, for example, we think of a group of patients in which type=1 and age is fixed , it makes little sense to think of this group of patients with varying duration. [age, and hence duration, are fixed]

Stepwise Methods

By stepwise methods, we discussing methods in which the choice of variables is carried out by an automatic procedure [algorithm]. Usually, the automated procedure takes the form of a sequence of tests with preassigned decision rules. These automatic procedures provide for statistical model selection in cases where there are a large number of potential models, and where the investigator has no clue [how proceed with the model selection]. There have been many techniques and criteria proposed over the years. One might be tempted to consider elaborate strategies based on hypothesis tests or so called “adjusted” R-square or the Akaike information criterion or various Bayesian information criterion or Mallows' Cp, or the false discovery rate or area under the curve... the list goes on. The 'false discovery rate' seems to be garnering attention these days.

A 'stepwise' algorithm may involve many 'stages' and may include:
a) Forward selection: which involves starting with no variables in the model or perhaps starting with a preassigned set of worthy variables, trying out the variables one by one and including them if the criteria above deems them 'worthy'.
b) Backward elimination, which involves starting with all candidate variables and considering them one by one based on the criteria and then deleting any that are not 'worthy'.

c) Methods that are a combination of forward selection and backward elimination, considering at each stage for variables to be included or excluded.

The algorithm stops 'searching' when the criteria used is deemed 'best'. Then a model (or a*short* list of models) is output along with the measure of the model's goodness.

The first widely used algorithm appears to have been proposed by Efroymson (1960).

Any method that "automates" the process of model construction is be viewed with cynicism. Criticisms of stepwise methods generate a lengthy and colourful list of articles/emails/blogs by very prestigious statisticians, biostatisticians and others. A brief set of highlights from some of these articles can be found at:

www.stata.com/support/faqs/stat/stepwise.html

...or for that matter, if you 'google' 'stepwise regression', you will get an avalanche of discussion about the problems and issues.

In the world of 'data mining', automated procedures have been returning to the attention of analysts. The whole topic of so called 'expert systems' can generate considerable debate.

The process of model construction is very time consuming, difficult and far from a simple set of rules and regulations. It is, perhaps, tempting to think that this very laborious and demanding step in research can somehow be passed over. This is, in part, due to the fact that, for many [novice] scientists, most of statistics is magic coming out of a very powerful 'black box' and that somehow such a black box must be better at model construction than, say, a clear thinking group of researchers agonizing and debating over the merits and demerits of a candidate model after extended time taken to review and interpret such a model's implications and then to consider another model (or models) and how such a model may better add to knowledge in the research area.

Gatekeeping

It was with the advent of fast computers in the 1960s, that regression analyses could be done with relative ease. Before computers, considerable effort was given to trying to find methods to bypass or minimize the calculations. Much of that effort can be studied with considerable advantage to get a clear understanding of the issues, but, alas, most (nearly all) of that work is not considered part of the mainstream anymore.

At the time, many statisticians claimed that regression analysis was being abused and misunderstood. [There was plenty of abuse... some would say there still is...] The statisticians were no longer the gate keepers of this 'technology'.

"Independent" Factors

The language of regression was in its infancy in some ways back in the 1960's. Some authors referred to the outcome variable as the 'dependent variable' in so far as the outcome was dependent [conditioned on] a collection of predictor variables. Unfortunately, at this time, some authors then referred to the predictors as the independent variables [because they weren't the dependent variable]. Many statisticians protested the use of 'independent' here and attempted to develop other namings like predictor variables or selector variables. There remains a considerable inertia to this day regarding this naming. It gets worse. The literature is now filled with phrases like 'independent factors' and/or

'independent predictors' and more and more muddle...

It would seem that 'most' of the time, when a researcher refers to the 'independent factors', they usually mean that such factors have been presented in an additive way (i.e. No interactions). However, there does not seem to be clear guidance on these matters and the cynical reviewer needs to dig deep these days to determine what is actually intended.

"Continuous"

Continuity has a precise mathematical definition. [have a look in your favourite calculus text]
Informally, a continuous variable is one for which, within the limits the variable ranges, any value is possible.
Age, weight, height and duration of illness are examples of continuous variables.
A 7 point "Likert" variable is not a continuous variable. The number of return visits during a study is not a continuous variable.
A variable that is not continuous is called "discrete".

The adjective "continuous" has crept into constant usage in regression analysis. Often, there is a decision to be made as to whether to use an actual variable as a predictor variable or to use a version of this variable with 2 or more levels based on cutoffs/thresholds. The real issue is whether the actual real variable affects the response in the linear way. If this is a plausible assumption, then such a use of the actual variable may be warranted. If the effect is not linear, then one option is to set up a set of indicator variables based on sensible thresholds and to then study the nature of the variable-response relationship. Unfortunately, authors now speak of the use of a 'continuous' variable if the actual values of a variable are used. The continuity of the variable is in fact irrelevant to the issue at hand. The real issue is the nature of the variable-response relationship. Indeed, it is certainly possible and reasonable that a predictor variable can clearly have only a discrete set of values and yet for the purposes of the assessment of conditional log odds has the linear effect on the response. Such a predictor variable can, then, with advantage, be included in the linear predictor even though the variable is most clearly not 'continuous'. It is far more helpful to refer to the possible linearity of a such variable rather than to merely to say it is 'in the model' as a continuous variable. The continuity or discreteness of a variable is relevant when such a variable is being considered as a 'dependent' variable however. More on this when we discuss linear regression and conditional means.

Logarithms

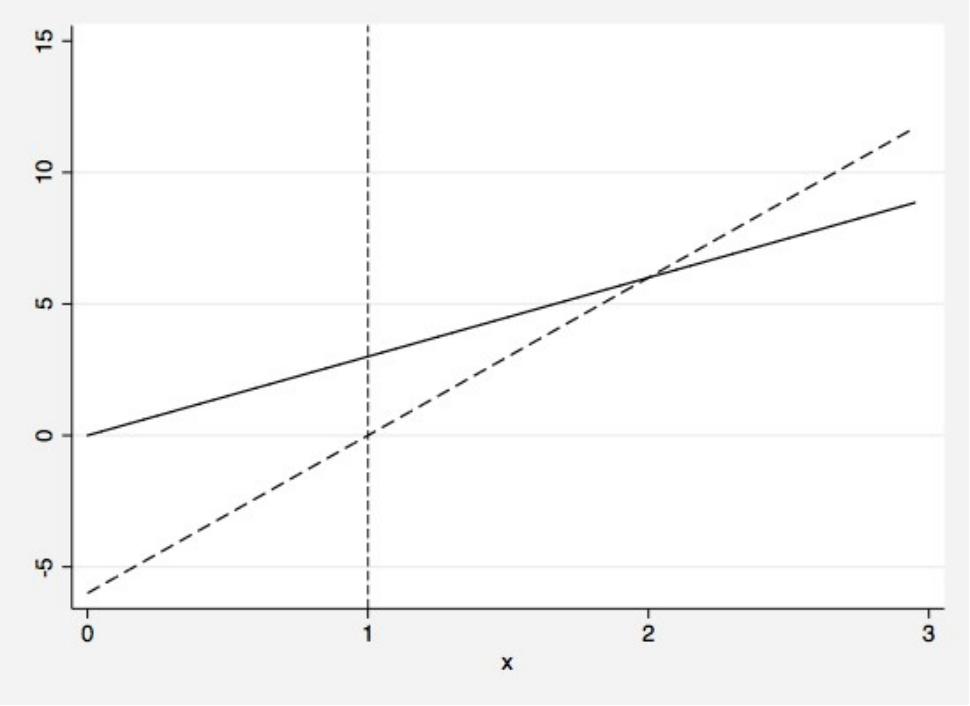
These days, all uses of logarithms are as 'base e' logarithms. It can be noted that such a choice of the base for logarithms has no real impact of any interpretation or description of log odds [or log anything] . Apart from a possible rescaling if, for example, an investigator wished to report logarithms to base 10, then all results would be rescaled by this same fixed quantity.

The use of the notation: $\ln(x) = \log_e(x)$ is not widely seen in epidemiology or biostatistics. For us, $\log(x)$ means $\log_e(x)$

$$\log_{10}(x) = \log_{10}(e)\log(x) = \frac{1}{\log(10)}\log(x) \approx 0.43\log(x) \quad \text{or} \quad \log(x) = \log(10)\log_{10}(x) \approx 2.3\log_{10}(x)$$

The Implications Of Centring

Consider a single variable x to be considered as the 'right hand side' of a regression model. A fit gives: $Y = b\,x$ A line through (0,0) with slope b.
If instead one considers x with a centred version x-c, then one gets a different fit: $Y = b'(x-c)$ A line through (c,0) with different slope b'.
The fitted values will be different for all x except when bx=b'(x-c) i.e when x= - b'c/(b-b')



The example above shows 2 lines: Y=3x and Y= 6(x-1). The first line is forced through (0,0) while the second line is forced through (1,0). The lines can have but one common fitted value at (2,6). So, with this simple model, centring changes the fitted values.

Now consider a model with both a constant 1 and the variable x. A fit gives $Y = b_0 1 + b_1 x$. Now consider the same model with 1 and x but again x is as centred x-c. Now the fit would be $Y = (b_0 + b_1 c)1 + b_1(x-c)$. This is the same line. Both versions give us the same set of fitted values.

a) Now, let us consider a model that may include many variables but the variable x appears in the model without any other forms of x such a quadratic or an interaction. If one wishes to ensure that the fitted values do not change when x is centred, then: $\beta\,x$ must be replaced by $\beta\,c\,1 + \beta(x-c)$ In other words, if x is to be replaced with x-c, then one must have the constant term 1 in the model.

b) Now suppose we are considering models that may include 2 variables x_1 and x_2 We wish to consider the centring of x_1 but we will not be centring x_2 . Now suppose we wish to include the term $x_1 x_2$. Then $\beta\,x_1 x_2$ must be replaced by $\beta\,c_1 x_2 + \beta(x_1 - c_1)x_2$ In other words, if x_1 is to be replaced with $x_1 - c_1$, then one must have x_2 in the model to ensure the fitted values are invariant to the centring. Notice that ensuring invariance here does not necessarily require either 1 or x_1 be included in the model.

c) Now suppose we consider the centring of both x_1 and x_2 and suppose we wish to include the term $x_1 x_2$. Then $\beta x_1 x_2$ must be replaced by $\beta c_1 c_2 1 - \beta c_2 x_1 - \beta c_1 x_2 + \beta (x_1 - c_1)(x_2 - c_2)$ In other words, if x_1 is to be replaced with $x_1 - c_1$ and x_2 is to be replaced with $x_2 - c_2$, then one must have 1 and x_1 and x_2 in the model to ensure the fitted values are invariant to the centring. This principle for the contents of such models goes by many names including 'well-formed' and ' hierarchically well formulated' . A related but not identical principle is often cited for the construction of analysis of variance tables.

d) Next, suppose that x_1 is to be centred but x_2 and x_3 are not to be centred and suppose we wish to include the term $x_1 x_2 x_3$. Then $\beta x_1 x_2 x_3$ must be replaced by $\beta c_1 x_2 x_3 + \beta (x_1 - c_1) x_2 x_3$ In other words, if x_1 is to be replaced with $x_1 - c_1$, then one must have $x_2 x_3$ in the model.

As a first 'real' example, consider dichotomous exposure E, gender G and actual age A. It makes no sense to consider the centring of the indicators E and G while age A could be conceptualized centred.

We could start from :

$\log p/((1-p)) = \beta_0 1 + \beta_1 G + \beta_2 A + \beta_3 GA + \beta_4 E + \beta_5 \mathrm{GE} + \beta_6 AE + \beta_7 GAE$

If we to include GAE, then above discussion requires the model to include GE. If we are to include GA we must include G. If we are to include AE, we must include E. If we are to include A, we must include '1'.

Reconsider this model and let us suppose that previous research indicates that, for the unexposed, the log odds of disease relationship with age does not depend on gender. This suggests the consideration of the model:

$\log p/((1-p)) = \beta_0 1 + \beta_1 G + \beta_2 A + \beta_4 E + \beta_5 \mathrm{GE} + \beta_6 AE + \beta_7 GAE$

If the previous research is reasonable here, we may have a clearer opportunity to see if age modification depends on gender and other forms of modification and then possibly confounding.

Further, if we decide to centre age A at $A = A_0$,say, we obtain the model:

$\log p/((1-p)) = \beta_0 1 + \beta_1 G + \beta_2 (A - A_0) + \beta_4 E + \beta_5 \mathrm{GE} + \beta_6 (A - A_0) E + \beta_7 G (A - A_0) E$

One needs to reinterpret the regression coefficients for '1', E, and GE as they are now specific to $A = A_0$. The fitted values will not change from the fitted values from the model with age A not centred. Indeed, re-expressing this last version gives:

$\log p/((1-p)) = (\beta_0 - \beta_2 A_0) 1 + \beta_1 G + \beta_2 A + (\beta_4 - \beta_6 A_0) E + (\beta_5 - \beta_7 A_0) \mathrm{GE} + \beta_6 AE + \beta_7 GAE$

With this writing, the coefficients for '1', E and GE are again specific to A=0

Notice that the consideration of the inclusion or exclusion of GA from the model need not based on the consideration of age centring, per se.

e) Now let us suppose that x_1 and x_2 are to be centred but x_3 is not to be centred and suppose again that we wish to include $x_1 x_2 x_3$. Then

$\beta x_1 x_2 x_3$ must be replaced by $\beta c_1 c_2 x_3 - \beta c_2 x_1 x_3 - \beta c_1 x_2 x_3 + \beta (x_1 - c_1)(x_2 - c_2) x_3$. So, here, we need x_3, $x_1 x_3$ and $x_2 x_3$ in the model.

f) If x_1, x_2 and x_3 are to be centred and $x_1 x_2 x_3$ is to be included, then:

$\beta x_1 x_2 x_3$ must be replaced by $-\beta c_1 c_2 c_3 +$ six more terms $+\beta (x_1 - c_1)(x_2 - c_2)(x_3 - c_3)$ and you can check that 1, $x_1, x_2, x_3, x_1 x_2$, $x_1 x_3$ and $x_2 x_3$ are all needed in the model. Another restatement of the 'hierarchically well formulated' [et al] principle.

g) One more... let us suppose that x_1 is to be centred and we wish to include x_1^2 in the model. Then: βx_1^2 must be replaced with $-\beta c_1^2+2\beta c_1 x_1+\beta(x_1-c_1)^2$ and so 1 and x_1 must in the model.

h) and on... There are many further extensions... cubics... more than one quadratic and so on...

There are, however, numerous circumstances in which centring would not be warranted. For example, the micro assay.... [next]

Micro Assay

Sometimes researchers will develop a study to compare 2 drugs [say]: a standard version and a test version. [E = 0 (standard) and E = 1 (test)] Both versions are being considered at very low dosages (D) and maybe a zero dose [or placebo] is also considered. In such a scenario, a starting point for analysis might involve an assumption of linearity at these low doses:

$$\log(p/(1-p))=\beta_0+\beta_1 E+\beta_2 D+\beta_3 ED$$

The assessment of β_1 might come first. This is sometimes called a 'validity' test because surely we need to have that a zero dose of the standard version of the drug is the same as a zero dose of the test version of the drug. If the validity test is not significant [or, if it is known that β_1 must be zero] then one considers a model like:

$$\log(p/(1-p))=\beta_0+\beta_2 D+\beta_3 ED$$

This model provides for 2 straight lines emanating from the same point β_0 . The lines have different slopes: β_2 and $\beta_2+\beta_3$ for the standard and test respectively.

Further, there may be advantage to recasting the model as:

$$\log(p/(1-p))=\beta_0+\beta_2 D \quad \text{for the standard}$$

$$\log(p/(1-p))=\beta_0+\beta_2 k D \quad \text{for the test}$$

or

$$\log(p/(1-p))=\beta_0+\beta_2 D+\beta_2(k-1) ED$$

With this recasting, k is called the relative potency of the test relative to the standard. A dose of x units of the test has an outcome that is the same as kx units of the standard. There are methods available for estimating the relative potency. Such methods use a result called Fieller's theorem. Further development of these techniques will take us too far afield.

Mutually Exclusive and Exhaustive Indicators

In our very first model based example, we considered:

p=Pr(E)

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 D$$

where D is an indicator for disease status. All participants are classified as have disease (D=1) or not having disease (D=0). The classification gives responses that are exhaustive. Everyone is either D=0 or D=1. The classification gives responses that are mutually exclusive. No one is both D=0 and D=1.

Now consider a case-control cancer study. E=1 (alternating) E=0 (sequential) Suppose participants are classified as:

R=1 (progression) R=2 (no change) R=3 (partial remission) or R=4 (complete remission). Each participant receives one and only one classification. Accordingly, R provides for mutually exclusive

and exhaustive options. Now let us define R_i as the indicator for R=i and consider the model:

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 R_1+\beta_3 R_3+\beta_4 R_4$$

it is important to note that $\sum_{i=1}^{4} R_i=1$ and further, the R_i are functionally related in that for a given participant if $R_1=1$, say, then the other 3 R_2, R_3, R_4 must be zero.
And if $R_2=1$ then the other 3 R_1, R_3, R_4 must be zero.
So, for participants with $R_1=1$ we have that:

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 1+\beta_3 0+\beta_4 0=\beta_0+\beta_1 1=\beta_0+\beta_1$$

And for participants with $R_2=1$ we have that:

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 0+\beta_3 0+\beta_4 0=\beta_0+\beta_1 0=\beta_0$$

Accordingly, we see that β_1 is the difference between the log of the odds of exposure for those with progression minus log of the odds of exposure for those with no change.
Similarly for β_3 and β_4
One must take care when considering the removal of any one term in a set of indicator variables as above. For example, if one considers $\beta_1=0$ and then assesses:

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_3 R_3+\beta_4 R_4$$

The baseline group is now those with either $R_1=1$ or $R_2=1$ and so, for example, we see that β_3 is now the difference between the log of the odds of exposure for those with partial remission minus log of the odds of exposure for those with either progression or no change.

Non Mutually Exclusive Indicators

When a set of indicators provides for a set of mutually exclusive and exhaustive groupings, we obtain a special form of interpretation of the associated coefficients. This arguably simple interpretation is not available when the indicators are not mutually exclusive. We have already seen many examples of this matter. Take for example a case-control study with age group (old A=1 young A=0) and a model like:

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 D+\beta_2 A$$

Here, β_1 is the 'assumed common' difference between the log of the odds of exposure for those with disease minus the log of the odds of exposure for those without disease. By 'assumed common' we mean that the difference applies to both the young and the old. Of course, if possible, the investigator would assess the assumption (that age is not a modifier) before consideration of the additive model.

It is important to be aware of the implications of the inclusion of non-mutually exclusive indicators in model assessment and interpretation. A rather extreme illustration should serve to make this issue clear.

Consider a database that contains an outcome of interest, say, myocardial infraction (MI) and an extended list of comorbidities. A real example came to my attention a while back that included more than 30 such comorbidities. For a start to this discussion, let us suppose the list was hypertension (H), diabetes (D), smoking (S) and obesity (O) and further we will suppose that each was coded as an

indicator (1=presence of the comorbidity; 0=absence of the comorbidity). Let us consider a model for p=Pr(MI).

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 H+\beta_2 D+\beta_3 S+\beta_4 O$$

In typical applications, Stata will carry out such a fit without objection. Such models might also include 'adjustment' for age and gender. We need not add age and gender to the mix to make the point to come. In any case, investigators may speak of the success of such models without regard to interpretation of the terms in such a model. A common error in interpretation would be to say that coefficients relate to a baseline group without such comorbidities.

To attempt an interpretation, take β_3 , for example. We must conceptualize 2 sets of individuals. Both sets have the same value for H, D and O. One set has S=1 and the other set has S=0. We then consider the difference in the log odds of MI between the set with S=1 and the set with S=0 and we require that this difference must be the same for each of the 8 combinations of H, D and O. Here 'assumed common' difference applies to all 8 of these combinations.

Imagine another model with all 30 comorbidities. Call them C_1 C_2 ... C_{30} and now imagine fitting:

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\sum_{i=1}^{30}\beta_i C_i$$

Again, Stata will fit such a model without objection (possibly deleting some terms... the least of our concerns here). Now the interpretation of any coefficient (say smoking: C_1) requires a conceptualization of 2 sets, one with smokers $C_1=1$ and one with nonsmokers $C_1=0$ and now there are 2^{29} combinations of the other 29 comorbidities. The 'assumed common' difference now applies to all 2^{29} pairs of 2 sets. Fantasy, indeed.

Such models have an illusion of simplicity in that some authors might think that the terms surely must have an simple (and realistic!) interpretation. Far from the case here.

Now, it is clear that even large databases cannot contain all of these combinations. So one might then think that since the 'assumed common' assumption cannot be assessed, then proceeding with such models has some scientific merit. Inevitably, the investigator is faced with a vastly more complex problem and simple expediency cannot be the driver.

One option of some potential would be to consider the commonly occurring sets of comorbidities and to construct a set of mutually exclusive combinations.

Models for Rate Ratios and Rate Differences

We have given most of our time so far to logistic regression. Modeling the log of the odds leads us to direct analogues with stratified analyses which are based on odds ratios. It can be argued that the poster child for logistic regression is the case-control study. But what of the [mighty] cohort study and of course, lest we forget the [gold coated] clinical trial. We have considered rate ratios and rate differences in standard ways via stratified analysis but we have ducked the option of models to handle rate ratios and rate differences. There are some good reasons why model based methods for rate ratios and rate differences have not reached the same attention as logistic regression.

Perhaps the biggest reason may not seem that important to epidemiologists but that biggest reason is that a log odds can be any number: positive or negative. No boundaries, as the mathematicians [and more importantly, the numerical analysts] say. We will see that modeling a rate ratio or a rate difference involves certain challenges we have not faced as of yet.

Whither Logistic Regression?

Sometimes we can compute Rate Differences or Rate Ratios from a Logistic Regression. When? Well, it depends...

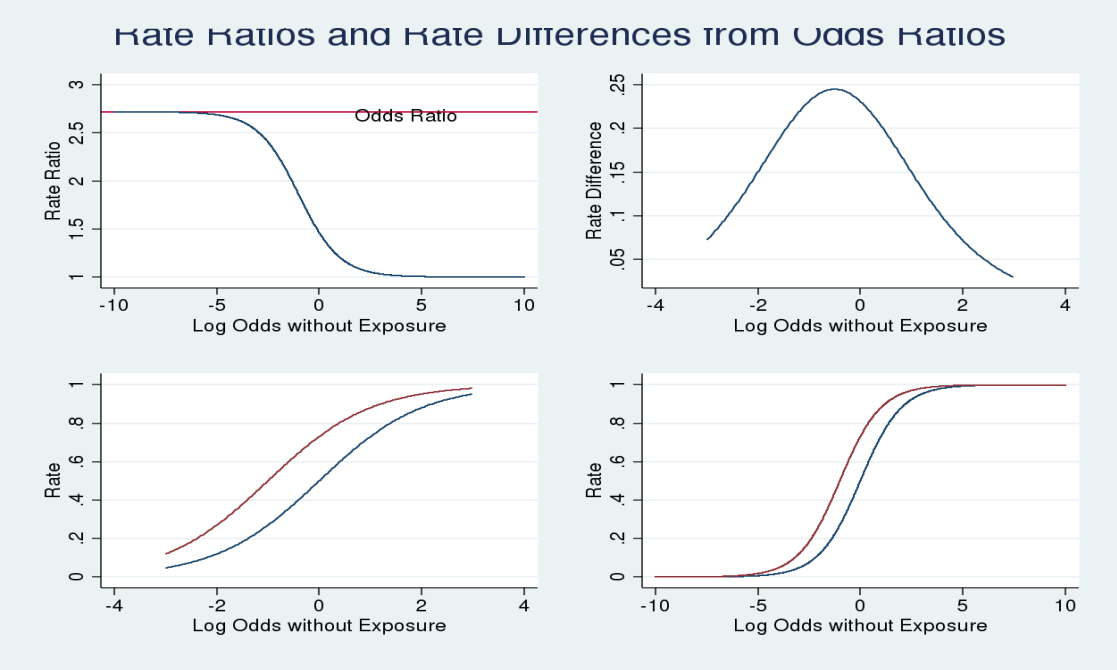
Lets look at an example to illustrate the challenges here.

Suppose we model the log odds using:

$$\log(p/(1-p)) = \beta_0 + \beta_1 E + \beta_2 A$$

and we get $\hat{\beta}_1 = 1$ so that the $\hat{OR} = 2.71$ What can we say about $\hat{RR}$ or $\hat{RD}$? It turns out that we can determine the $\hat{RR}$ or the $\hat{RD}$ once we know the log odds with exposure.

The graphs below illustrate the issues:



From the upper left hand graph, we can see that the rate ratio estimate and the odds ratio estimate are the same when the log odds of disease without exposure is "small". (log odds of disease in the absence of exposure less than -5)

The rate difference estimate depends on the log odds of disease in the absence of exposure and in more complex ways. For log odds of disease in the absence of exposure between -2 and 2, the rate difference estimate is varies between about 0.1 and 0.25

For this illustration at least, we see that if we wish to make inferences about RD or RR we get useful information from logistic regression in very limited settings.

We are then directed to the direct modeling methods.

Log-Binomial Regression

Let p=Pr(Disease) and consider: $\log(p) = \sum \beta_i x_i$

As an example, let us return to the NASCET project with p=Pr(Stroke), D: stroke E: stenosis group (Elevated=1; Not elevated=0), Age Group (Young=0; Old=1) and Gender (F=0; F=1) as potential confounders/modifiers. As an illustration only, consider:

$$\log(p) = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AG + \beta_5 EA + \beta_6 EG + \beta_7 EAG$$

then all of the coefficients are now interpreted in terms of log of probability of disease [here; log of risk of stroke] For example, β_1 is, for the young females, the log of the risk of stroke for those with

elevated stenosis minus the log of the risk of stroke for those without elevated stenosis. Just like before, using that fact that the exponent of a difference is the ratios of the exponents, then:

e^{β_1} = the relative risk for the young females

As so we can have exponents of coefficients yielding rate ratios and, like before, we can have ratios of rate ratios [and ratios of ratios of ratios....]

Identity-Binomial Regression:
Now we have: $p = \sum \beta_i x_i$
Now continuing with the last example, consider:
$p = \beta_0 + \beta_1 E + \beta_2 A + \beta_3 G + \beta_4 AG + \beta_5 EA + \beta_6 EG + \beta_7 EAG$
For this model, β_1 is now the risk difference for young females. No exponentiating here.
So the coefficients here will be risk differences or differences between risk differences [and differences of differences of differences....]

In principle, models based on the log link will reproduce stratified analysis components based on rate ratios while models based on the identity link will reproduce stratified analysis components based on rate differences. The same qualifications as with logistic regression apply here since such models are fit via likelihood methodology while the Mantel-Haentzel methodology has slightly different approximations in their development.

The big catch [22?] with these models is the inherent boundaries of log(p) and p. Probabilities (p) must be between 0 and 1 and so all the fitting of rate differences must obey this "constraint". The same matter applies to log(p) which must be negative. With "large" sample samples sizes and fitting algorithms carried out away from boundaries, these constraints have little impact but with "modest" studies and with the 'inevitable' [good thing!] small probabilities/rates/risk, the algorithms can bump into boundaries and then the "search" for a maximum [of a likelihood] can fail. This matter has been receiving serious attention [notably TW, ME & GHF(2014) and GS & GHF(2019)_] .

In the last few years, major strides have been made with log-binomial models. Such advances have been implemented in R:

https://cran.r-project.org/package=lbreg

Such advances have not [yet] been implemented in Stata. One can try binreg in Stata but there can be serious problems.
[from 'help binreg' in Stata]

```
binreg fits generalized linear models for the binomial family.  It
    estimates odds ratios, risk ratios, health ratios, and risk differences.
    The available links are

                Option    Implied link               Parameter
                ------------------------------------------------
                or               logit     odds ratios = exp(b)
                rr                 log     risk ratios = exp(b)
                hr      log complement   health ratios = exp(b)
                rd            identity     risk differences = b

    Note that estimates of odds, risk, and health ratios are obtained by
    exponentiating the appropriate coefficients.  The option or produces the
    same results as Stata's logistic command, and or coefficients yields the
    same results as the logit command.  When no link is specified or implied,
```

```
    or is assumed.
```

The 'link' g is a function of the probability – p. Generally, then,

$$g(p)=\beta_0+\beta_1 E+\beta_2 A+\beta_3 G+\beta_4 AG+\beta_5 EA+\beta_6 EG+\beta_7 EAG$$

when g(p) = log(p/(1-p)), we have binomial regression with a logit link (logistic regression)
when g(p)= log(p), we have binomial regression with a log link (rate ratio regression)
and when g(p)=p, we have binomial regression with an identity link (rate difference regression).

Anytime Stata gives an 'note' or a 'warning' message, you need to take heed. The matters leading to notes and warnings can occur more often with the (non-logit link based) binomial regressions.

Lets take a brief look at a part of a NASCET dataset (courtesy M.E.) and a start at an analysis based on risk ratios (risk of stroke for those with elevated stenosis over the risk of stroke for those without elevated stenosis)

```
. gen sten=(stengrp>1)

. gen stro=stroke-1

. egen genage=group(sex agegp)

. cs stro sten,by(sex agegp)

       sex agegp |       RR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
             1 1 |    .4720497     .1493356   1.492149     3.833333
             1 2 |    .8939394     .2683331   2.978118     2.563107
             1 3 |    1.403509     .3951443   4.985108     1.628571
             2 1 |    1.468421     .7497855   2.875837      6.06383
             2 2 |    2.509804     1.362044   4.624751     6.181818
             2 3 |    1.501235     .7506118   3.002491     4.879518
-----------------+-------------------------------------------------
           Crude |    1.559672     1.126613   2.159194
    M-H combined |    1.516141     1.095846   2.097634
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(5) =    7.325  Pr>chi2 = 0.1976

. cs stro sten,by(genage)

group(sex agegp) |       RR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               1 |    .4720497     .1493356   1.492149     3.833333
               2 |    .8939394     .2683331   2.978118     2.563107
               3 |    1.403509     .3951443   4.985108     1.628571
               4 |    1.468421     .7497855   2.875837      6.06383
               5 |    2.509804     1.362044   4.624751     6.181818
               6 |    1.501235     .7506118   3.002491     4.879518
-----------------+-------------------------------------------------
           Crude |    1.559672     1.126613   2.159194
    M-H combined |    1.516141     1.095846   2.097634
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(5) =    7.325  Pr>chi2 = 0.1976
```

$$\log(p)=\beta_0+\beta_1 E+\beta_2 G_2+\beta_3 G_3+\beta_4 G_4+\beta_5 G_5+\beta_6 G_6+\beta_7 EG_2+\beta_8 EG_3+\beta_9 EG_4+\beta_{10} EG_5+\beta_{11} EG_6$$

```
binreg stro i.sten#i.genage,rr
i.sten            _Isten_0-1          (naturally coded; _Isten_0 omitted)
i.genage          _Igenage_1-6        (naturally coded; _Igenage_1 omitted)
i.sten*i.genage   _IsteXgen_#_#       (coded as above)

Generalized linear models                         No. of obs      =       724
Optimization     : MQL Fisher scoring             Residual df     =       712
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  645.2029024                   (1/df) Deviance =  .9061839
Pearson          =  723.9978177                   (1/df) Pearson  =  1.016851

Variance function: V(u) = u*(1-u)                 [Bernoulli]
```

```
Link function    : g(u) = ln(u)                  [Log]

                                                 BIC             = -4043.169

------------------------------------------------------------------------------
             |                 EIM
        stro | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Isten_1 |   .4720497   .2771857    -1.28   0.201     .1493364    1.492141
  _Igenage_2 |   .5520581   .2848628    -1.15   0.250     .2007999    1.517771
  _Igenage_3 |   1.017857   .6335083     0.03   0.977     .3005412    3.447225
  _Igenage_4 |   .7004608   .3046172    -0.82   0.413     .2986842    1.642689
  _Igenage_5 |   .5816327   .2541579    -1.24   0.215     .2469998    1.369623
  _Igenage_6 |   1.285714    .576907     0.56   0.575     .5335876    3.098013
_IsteXgen_~2 |    1.89374   1.608883     0.75   0.452     .3582258    10.01115
_IsteXgen_~3 |   2.973223   2.597107     1.25   0.212     .5366655    16.47218
_IsteXgen_~4 |   3.110734   2.115319     1.67   0.095     .8204279    11.79466
_IsteXgen_~5 |   5.316821   3.534987     2.51   0.012     1.444492    19.56992
_IsteXgen_~6 |   3.180247   2.179971     1.69   0.091     .8298237     12.1881
------------------------------------------------------------------------------
```

Some parts of this output use slightly different namings. Deviance is -2*log-likelihood. So that the iteration sequence showing deviance going down is analogous to log-likelihood going up.

The red highlighted rows in the above table show the estimated risk ratio for young females of 0.4720 as obtained from the stratified analysis. The number 5.3168 is in fact an estimated ratio of risk ratios. The we get that the estimated RR for middle aged males is 0.4720*5.3168 = 2.5098 which the estimated RR for middle aged males in the stratified analysis.

Here is a similar analysis of risk differences:

$$p=\beta_0+\beta_1 E+\beta_2 G_2+\beta_3 G_3+\beta_4 G_4+\beta_5 G_5+\beta_6 G_6+\beta_7 EG_2+\beta_8 EG_3+\beta_9 EG_4+\beta_{10} EG_5+\beta_{11} EG_6$$

```
binreg stro i.sten#i.genage,rd
i.sten            _Isten_0-1          (naturally coded; _Isten_0 omitted)
i.genage          _Igenage_1-6        (naturally coded; _Igenage_1 omitted)
i.sten*i.genage   _IsteXgen_#_#       (coded as above)

Generalized linear models                         No. of obs      =       724
Optimization     : MQL Fisher scoring             Residual df     =       712
                   (IRLS EIM)                     Scale parameter =         1
Deviance         =  645.2029024                   (1/df) Deviance =  .9061839
Pearson          =          724                   (1/df) Pearson  =  1.016854

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = u                       [Identity]

                                                  BIC             = -4043.169

------------------------------------------------------------------------------
             |                 EIM
        stro | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Isten_1 |   -.097254     .07537    -1.29   0.197    -.2449765    .0504685
  _Igenage_2 |  -.0825156   .0741823    -1.11   0.266    -.2279103     .062879
  _Igenage_3 |   .0032895   .1160868     0.03   0.977    -.2242365    .2308154
  _Igenage_4 |  -.0551783   .0718546    -0.77   0.443    -.1960106    .0856541
  _Igenage_5 |  -.0770677   .0693455    -1.11   0.266    -.2129823     .058847
  _Igenage_6 |   .0526316   .0933337     0.56   0.573    -.1302991    .2355623
_IsteXgen_~2 |   .0864682   .0954321     0.91   0.365    -.1005753    .2735116
_IsteXgen_~3 |   .1729119   .1593979     1.08   0.278    -.1395023    .4853261
_IsteXgen_~4 |   .1576954   .0922259     1.71   0.087     -.023064    .3384549
_IsteXgen_~5 |   .2590187   .0904812     2.86   0.004     .0816789    .4363586
_IsteXgen_~6 |   .2159675   .1246157     1.73   0.083    -.0282748    .4602097
       _cons |   .1842105   .0628861     2.93   0.003     .0609561     .307465
------------------------------------------------------------------------------
```

Notice that we are seeing a very similar finding. The estimated risk difference for middle aged males is: -0.0972+0.2590= 0.1617

which is the same as the stratified analysis:

```
. cs stro sten if genage==5

                 | sten                    |
                 |   Exposed   Unexposed   |      Total
-----------------+-------------------------+------------
           Cases |        32          12   |         44
        Noncases |        87         100   |        187
-----------------+-------------------------+------------
           Total |       119         112   |        231
                 |                         |
            Risk |  .2689076    .1071429   |   .1904762
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
 Risk difference |          .1617647       |    .0636449    .2598845
      Risk ratio |          2.509804       |    1.362044    4.624751
 Attr. frac. ex. |          .6015625       |    .2658095    .7837721
 Attr. frac. pop |             .4375       |
                 +-------------------------------------------------
                               chi2(1) =     9.79  Pr>chi2 = 0.0018

. disp -0.097254+0.2590187
.1617647
```