AGENDA

Key Points in / Commentary on ALR Ch 4 —

- General comments on purposes of multivariable models
- Variable Selection [4.2]
  - "univariate" analyses, and what they can sometimes miss
  - parsimony, confounding, p-values, overfitting
  - H&L's 5 steps.
    - nominal/ordinal X's
    - continuous X's .. 'linear first, refine later' (with caveats)
    - comments on 'interactions'
    - smoothing & other ways to decide ***form*** of continuous X
    - partial regression (leverage) plots
- Stepwise[4.3]
- Best Subsets [4.4]

Classification & Regression Trees [not in ALR]

Note: Although ALR Chapter 4 has mostly 'generic' material that is common to all types of regressions,and Chapter 5 is more logistic-regression specific, it is difficult to skip right to Ch 5 without dealing with the main points in ALR Ch 4.

Readings (* primary)

* Chapter 4 of Hosmer and Lemeshow

Other Resources

texts

HARRELL, Frank E. Regression Modeling Strategies. With applications to Linear Models, Logistic Regression, and Survival Analysis. Springer Verlag 2002.

articles

- Modeling & Variable Selection in Epidemiology GREENLAND, S. American Journal of Public Health 1989 [on website... via alr_4]
- Multivariable prognostic models: issues in developing models. evaluating assumptions and adequacy, and measuring and reducing errors by Frank E HARRELL, Kerry L LEE and Daniel B MARK , Statistics in Medicine 1996 [on website... via alr_4].. superseded by Harrell's 2002 textbook

Other Links

- to material on Classification and Regression Trees [on website... via alr_4]

Key Points in / Commentary on ALR Ch 4 ––

- **" perceived to be only one possible model" *(4.1 page 91, para 1)***

  All models are wrong. Some models are useful.[1]

  George Box, Statistician & Author, Univ. of Wisconsin [of Box-Cox fame]

- **" 'best' model within the scientific context of the problem" *(para 2)***

  Scientific context / focus may be on:-

  - 'etiologic' variable, but need adjustment for confounding by others [fairer]
  - etiologic variable; sharpen estimate if remove noise from others [sharper]
  - etiologic variable, but modification of its effect by others [specificity]
  - diagnostic/prognostic function of several variables [accuracy, stability, ...]

  The approach should reflect the context/purpose. For example, in an etiologic mode, the primary purpose is control of confounding, and less on the "best" model. Collinearity of two X variables is a serious issue only if the focus is on one or both of these two variables; clinical prediction models do not necessarily focus on individual variables, so much as on the score used to compute probability = exp[score]/(1+this). If the score is used in something like the Framingham Risk Score, then it can suggest to the person or healthcare professional who computes it which items in the score are adding most to the risk.

- **"must have " *(para 2)***

  - basic plan for selecting the variables for the model

  - a set of methods for assessing the adequacy of the model both in terms of its individual variables and its overall fit.

  As we will see, the order in which things "should" be done (1. selecting variables, and 2. making sure variables are represented in the "best" ***form*** or scale e.g., linear, categorical, quadratic, threshold, etc.., is difficult to prescribe. In a multi-dimensional situation, it is not easy to deal with all issues, or visualize all variables, at once: using one form for a variable at a selection step, hoping to 'fix up' the form, or not considering a biologically relevant effect modifier, until a later stage, i.e. after the number of variables has already been reduced, may mislead. ***Plots*** of the data are very important.

  Decisions made only on the basis of so-called 'univariate' analyses can be misleading.. after all, if one 'X' variable 'confounds the effect of another, then the results of the separate univariate analyses will be misleading (remember that confounding can ***mask*** effects).

  Some textbook authors take -- to JH at least -- an overly suspicious attitude to Nature, and suggest starting with the largest possible models, including interaction (product) terms for all pairs of variables.

  Data-analysis is both art and science, or rather ***substance*** as well as computation. If it were strictly a question of finding the "best fitting model to describe relation ship between Y and a set of X variables", we could simply label our 20 variables X1-X20. (I have seen students from statistics departments do this, presumably to hoping to save time typing, and wear and tear on their fingernails). We could then ship the data off to a computing robot, who could follow some "stepwise regression" recipe and give us back an equation linking Y and a set of X variables.

  It is one thing to have 'information' (documentation, recollections? facts?) on each individual's smoking 'history'. It is quite another to turn that set of information into ***terms*** in the regression i.e., to ***design*** the form of the regression terms. Miettinen, in his Theoretical Epidemiology text, devotes considerable space to this important data-analysis activity, and you might appreciate his structured approach when you have to face this activity in your thesis work.

- 4.2 **VARIABLE SELECTION *(page 92, para 1)***

  "traditional approach: ***parsimony*** / Occam's razor[2]
  ..keep number of variables to minimum:- numerically stable model
  ..the more variables included, the greater the estimated standard errors"

  This last part about increased SE's is much more critical with ***logistic*** regression i.e., with the ' y|x ~ Bernoulli(some function of Bx)' models, than with ' y|x ~ Normal(some function of Bx, σ)' models.

---

[1]*from Stephen Duffull, School of Pharmacy, Univ of Queensland: "I have been trying to find the original reference for the famous quote: "All models are wrong but some are useful" GEP Box. After searching around, I found several hits for this - most were wrong but some were useful. It seems that this statement is made in a number of forms in various articles written by Box (although only 1 matching the exact wording given above). The only version I have found to date is in: Box GEP Robustness in the strategy of scientific model building. In: Launer RL & Wilkinson GN Robustness in Statistics. New York: Academic Press, 1979:pp. 202.

Someone else (found via web search) spoke of paraphrasing George Orwell that, "All models are wrong, but some models are less wrong than others."

[2][from http://en.wikipedia.org/wiki/Occam's_Razor] The principle is most often expressed as Entia non sunt multiplicanda praeter necessitatem, or "Entities should not be multiplied beyond necessity", but this sentence was written by later authors and cannot be found in Occam's surviving writings. William wrote, in Latin, Pluralitas non est ponenda sine neccesitate, which translates literally into English as "Plurality should not be posited without necessity".

Dave Beckett of the University of Kent at Canterbury writes: "The medieval rule of parsimony, or principle of economy, frequently used by Ockham came to be known as Ockham's razor." [1]

Occam's Razor has also been referred to as "parsimony of postulates" and the "principle of simplicity" and "K.I.S.S." (keep it simple, stupid). Another proverb expressing the idea that is often heard in medical schools is, "When you hear hoofbeats, think horses, not zebras."

"epidemiologic methodologists suggest including all clinically and intuitively relevant variables in the model, regardless of their "statistical significance"

This is not an entirely accurate statement of prevailing dogma. Yes, epidemiologists do suggest a variant of this in the control of confounding, but realistically use a "change in estimate" rule as well, since they realize that one cannot include ***all*** such variables in the model. And yes, they do, as they should, advise against using statistical significance as the criterion for confounding. JH's approach, especially in looking at imbalances in Table 1 of a trial, or any other comparison, is to look for "embarrassing differences", or to say to a physician: if you had to take care of the patients in column 1, and your colleague those in column 2, which of you would have the tougher task? JH also reinforces the advice in the RCT methodology literature that one should not put p-values beside the differences with respect to each baseline variable (row), even if the NEJM still allows authors to do so.

"The major problem with this approach is "overfitting"

The 'excellent tutorial paper" by Harrell et al can be found on the c681 webpage (under alr_4), along with the equally helpful one by Greenland 1996.

Think of 'overfitting' as being overly particularistic and of "tuning" the model to the data. A radio that is "tuned" to bring in as specific station perfectly when operated at the corner of Peel and Pine maybe not receive the signal very well at another location.

- **4.2 SEVERAL STEPS TO AID IN SELECTION OF VARIABLES (p92-)**

Because, as these authors state, the "process is quite similar to the one used in linear regression" (they mean in c621), the comments below will try to focus on what (little) is specific to logistic regression.

- **1 (...) begin with careful univariate analysis of each [X] variable *(p 92)***

I am happy to see that this textbook does not start with examining the distribution of the X variable , and "check it for Normality". Over the years, I have found that students obsess with checking normally of X's, when in fact the variation of interest is in the Y's, CONDITIONAL on the X values. Indeed, as I say elsewhere in c697 and v678, it is better, for the precision of slopes, etc., NOT to have X's be Normal (Gaussian). And after all, if X1=age and X2= sex, why check that age is Gaussian? Why not check (and confirm) that sex (i.e. the random variable with 2 categories, male-female -- sometimes, and correctly in non-biological, i.e. more 'sociological' situations, called 'gender') ***cannot*** have a Gaussian distribution.

Of course, 'wild' values of an X variable, measured on a numerical scale, can have a strong influence on the fitted coefficients, particularly in the case of logistic regression. And so one needs to watch out for these 'potentially influential' values.

For a 2 x k table, the usual test of homogeneity of proportions is a chi-square test. Since you are going to be using logit models anyway, you can test via the likelihood ratio test of (the k-1 beta_s that accompany the k-1 indicator (dummy) variables are zero, i.e. a model with a single beta_0 (a 'single proportion fits all' model) adequately explains the observed variation among the k proportions. One thing to be aware of: the chi-square test with k-1 df is an omnibus ***non-specific test***, and is not very sensitive to monotonic patterns in the proportions.. a single df test for trend in the proportions might more sense, especially if on expects a monotonic relationship. This illustrates why one cannot 'screen' variables of an ordinal nature with global (df>1) tests: if we make our selection based on p-values, without actually plotting the proportions, we might not keep the variable in the model, and so later on (in step 4 on p97), this variable wouldn't be there to have its 'form' scrutinized.

- **(...) for continuous variables, the most desirable univariate analysis involves fitting a univariate logistic regression model to obtain the estimated regression coefficient, the LR test, the SE and the Wald test" *(p 94)***

It is a important that this seemingly blanket statement is modified in the next paragraph with a "supplementary evaluation", to check if the logit is indeed linear in the X in question (H&L refer to the "appropriate scale".) And, of course the scientific question may focus specifically on the shape of the curve, as in the risks of spontaneous abortion as a function of folate levels -- as in the study recently reviewed in the practicum.

The "***smoothing***" referred to on page 94 can be carried out with the klm [lowess in version 8] command in Stata, or via (among others) the "FIT" command inside the interactive data menu in SAS (the "INSIGHT" module) -- more in 4.4.

This issue of how to represent a continuous variable at the early stages is an important, and slightly tricky, one. (With the caveat 2 paragraphs below) JH agrees that it makes more sense and is more relevant to deal with the correct form of a continuous X variable in the context of the multivariable model, i.e. with other relevant variables already in the model. If one devotes a lot of effort to its form at the univariate stage, the shape of the X<>Y relationship will be affected by all of the artifacts that can occur when we look at X<>Y relationships one X at a time.

- **2 (...) select variables for multivariate analysis *(p 95)***

Missing from the advice is any consideration of the possibility that the different data items may be obtained in 'blocks' (e.g., history, physical, blood tests, imaging, etc.) or that different items may cost more (in money, or discomfort..) to obtain.

The warning (in the 3rd para of (2)) about effects being 'masked' by confounding is an important one. For example, if we go back to the altitude in relation to east/west and north/south, we could see a situation where the specific pattern of (x1,x2) data points could lead us to conclude-- in univariate analyses -- that neither X matters, and stop right there (imagine the data points included were those encountered by travelling one block south for every 4 blocks east.!)

Whereas it may be possible in $y|x \sim N(*,\sigma)$ data-analyses to "***begin the multivariable analysis with all possible variables***", it seldom is with logistic

regression, where the "***effective" sample size***" is the ***smaller*** of the number of observations with Y=1, or with Y=0. [see sample size issues in Chapter 8.5]

Moreover, whether in y|x~N(*,σ) or in y|x~Bernoulli(*), there is another factor -- how correlated the X's are. If, conditional on X1, the distribution of X2 has a limited range, then the effective size for studying the net effect of X2 is reduced by (1 minus the square of the correlation between X2 and X1). You may have met this factor as the 'tolerance', or the reciprocal of it, i.e. 1/(1 minus the r-square between one X and all the other X's) as the Variance Inflation Factor (VIF).

"The analyst, not the computer, is ultimately responsible for the review and evaluation of the model" [end of (2)]

Here here! d'accord!

- **3 (...) Following the fit of the multivariate model, the importance of each variable included in the model should be verified *(p 96)***

JH has nothing to add

- **4 (...) Once we have obtained a model that contains the essential variables, we should look more closely at the variables in the model. *(p 96)***

"The question of appropriate categories for ***discrete*** variables should have been discussed at the univariate stage" (i.e. (1) |

"**For continuous variables, we should check the assumption of the linearity of the logit"**

This, presumably is the same strategy that H&L would recommend for a y|x~N(*,σ) situation.

Their logic that one can represent a continuous X as a linear effect in the preliminary stapes, and that it can be refined at step (4) is reasonable, provided, as they also warn, that the relationship is not U-shaped and missed entirely at step (0). This is where graphical displays, and some anticipation, with guidance from the literature or experts, are critical.

5(...) Once we have refined the main effects model and ascertained that the continuous variables are scaled correctly (see later), we check the interactions among the variables in the model *(p 98)*

- Note the sensible advice about having a ***biological*** basis for these "interactions". Indeed, these should be entertained a-priori. Sub-group analyses, based on data-dredging", run the risk of false positive findings.

- As it happens, the ***power to detect interactions is much less than the power to detect main effects***. This is more evident if one regards the beta_hat for the product term not as just another regression coefficient, but as a difference of two coefficients (slopes), one from the observations where the modifier M=0 and one from those where M=1. A simple example, from c621 type data, will illustrate:

suppose that we have n observations in all, n/2 with M=0 and n/2 with M=1. Suppose further that for each of the 2 levels of M, one half or n/4 have X=0, and n/4 have X=1. Then, within each level of M, and with the variance of slope comparing the average response when X=1 with that when X=0 is $\sigma^2 \times (1/ [n/4] + 1/[n/4])$. Thus the variance for the difference, in the Y<->X slopes, between the 2 levels of M is

$$\sigma^2 \times (1/ [n/4] + 1/\{n/4\}) + \sigma^2 \times (1/ [n/4] + 1/\{n/4\})$$

or

$$\sigma^2 \times (16/ n)$$

whereas, if the data are combined (to give n/2 with X=0 and n/2 with X=1, and there is no interaction, the variance of the estimated (common) slope is only

$$\sigma^2 \times (1/ [n/2] + 1/\{n/2\})$$

or

$$\sigma^2 \times (4/ n)$$

i.e. the variance for the estimate of the INTERACTION is 4 TIMES LARGER!

The same type of calculations applies for odds ratios: $V_0 = 1/a_0 + 1/b_0 + 1/c_0 + 1/d_0$ is the variance of the log odds ratio in the M=0 subgroup, and $V_1 = 1/a_1 + 1/b_1 + 1/c_1 + 1/d_1$ for those with M=1. Then the variance of the ***difference*** of two log odds ratios (the coefficient that accompanies the X•M product term), is $V_0 + V_1$, is also approximately 4 times larger than the variance of a single log odds ratio computed from the overall total of n observations, i.e.. V = 1/a + 1/b + 1/c + 1/d.

- when dealing with interaction terms, coefficients will be more manageable, and interpretable, and less correlated, if you center the components of the product before making the product. The example from the Lidkopping injury prevention study is a good case in point.

- Note the sensible advice about having a ***biological*** basis for these "interactions". Indeed, these should be entertained a-priori. Sub-group analyses, based on data-dredging", run the risk of false positive findings.

- **4 (...) we should look more closely at the variables in the model. ...** REVISITED on pages 99 onwards; this is the main focus of the example that starts on page 104.

- The "bottom line" advice is that given in the second half of page 99: ***Instead*** of categorizing the continuous X into quartiles and looking at the relationship with Y in a ***univariate analysis***, do so more in the (more realistic) multivariate analysis.

  So do not form and plot the empirical logits (univariate) that some other texts recommend. Instead, fit the other covariates plus the 3 indicator (design) variables for the quartiles of the X variable under scrutiny.

All JH would add to the advice is that one take account of the CI's that accompany the 3 beta_hats: without these, users tend to over-interpret apparent non-linearities -- which may be merely ***random*** fluctuations.

- The reason why the authors need to resort to "smoothers" to see the pattern has to do with the ***binary*** nature of Y, the same reason that in Fig 1.1 in Chapter 1, one cannot not see the forest from the trees until one computes the proportions where Y=1 for various groupings or categories of the "X" variable. Think of all of these smoothers (even the simplest one in Table 1.2 and Fig 1.2) as variations on the same theme.. some smoothers use "moving" averages, where they move or slide the "X" window; others (e.g. Fig 1.2) take non-overlapping 'windows' or 'slices' of the X axis.
- In c621, this issue of trying to see the correct form for a continuous X is made easier by the use of "partial leverage plots". They may go by various names in different packages, but the idea is basically this. One might be tempted to (1) compute the Y residuals from a model with the other X variables, and then (2) plot these residuals from model (1) against the X variable under scrutiny. But this would give a distorted view, since -- to the extent to which the "X" in question is correlated with (can be predicted from) the other X variables -- X has already been (partially) 'used up' in fitting model (1) and computing these residuals. For a good example of this, and of how to ***correctly*** view a multiple regression as a series of univariate regressions, look at the example on the next column. The key is to regress the residuals from (1) NOT on the X of interest, BUT on what remains by way of variation in X, after that variation in X which could have been predicted from the others is ***removed***. I have put an example of this, using the (low) birthweight data set, on the resources for alr_4 web page.
  This concept is at the heart of 'partial leverage plots', available in the interactive SAS INSIGHT or '**partial regression plots**' available by specifying the PARTIAL option in the MODEL statement in SAS PROC REG. In Stata, you can do this with the post-estimation command 'avplot indepvar' to graph an added-variable plot (leverage plot) after regress
- Read through the example on pp. 104-, paying attention to the smoothers used in diagrams Fig4.2 to 4.6 These strategies are best practiced in the context of the full-blown data-analysis project i c621 for example, or with the book beside you when dealing with your thesis data.

• 4.3 Stepwise and 4.4 (Best Subsets) Logistic Regression *(pp 116-)*

Not much to add, over what will be covered generically, in c621. Generally, other than for "pure" prediction situations, e.g., diagnostic and prognostic functions, we should use these 'partially mindless' approaches sparingly.

Multiple regression as a sequence of simple regressions

E.g.: Regression of Weight(lb) on age(yrs) and Height(in) in 11-16 year olds

3 SIMPLE REGRESSIONS

(1) WEIGHT = -105.378 + 3.363 * HEIGHT + RESWT

(2) AGE = -0.789 + 0.226 * HEIGHT + RESAGE , so that

(2') RESAGE = AGE - { -0.789 + 0.226 * HEIGHT }

(3) RESWT = -0.023 + 2.822 * RESAGE + RESIDUAL (variance 187.02)

Substitute (2') into (3) to get

(4) RESWT = -0.02337 + 2.822 * { AGE - {-0.789 + 0.226 * HEIGHT } }

and then (4) into (1) to get ...

(5) WEIGHT = -105.378 + 3.363 * HEIGHT +
-0.023 + 2.822 * { AGE - {-0.789 + 0.226 * HEIGHT } }

+ RESIDUAL (variance 187.02)

= -105.378 + 3.363 * HEIGHT + 2.822 * AGE
-0.023 + -2.822 * 0.226 * HEIGHT +
2.822 * {-{-0.789}}

+ RESIDUAL (variance 187.02)

= -103.174 + 2.725 * HEIGHT + 2.822 * AGE

+ RESIDUAL (variance 187.02)

This is numerically equivalent (apart from some rounding errors introduced by not using enough decimal places) to performing a multiple linear regression:

DEP VAR: WEIGHT N: 233 MULTIPLE R: 0.703 SQUARED MULTIPLE R: 0.494
ADJUSTED SQUARED MULTIPLE R: 0.490 STANDARD ERROR OF ESTIMATE: 13.70530

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	-103.14981	14.19908	0.00000	.	-7.26454	0.00000
HEIGHT	2.72315	0.29462	0.55333	0.61379	9.24288	0.00000
AGE	2.82220	0.80680	0.20941	0.61379	3.49802	0.00056

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	42186.19420	2	21093.09710	112.29578	0.00000
RESIDUAL	43202.08906	230	187.83517		

Classification and Regression Trees (CART)

One interesting approach that is not mentioned in the text is the use of '(regression) trees'. The idea is very appealing, and more natural that the 'purely mathematical' approach of linear predictors. I have put a link to a survival analysis example in the alr_4 resources page. A flavour for this method is given by the following abstract, from an author who has used this (also known as "Recursive partitioning") method extensively.Note also that he tests the algorithm in new ("validation") samples.

Goldman L, Weinberg M, Weisberg M, Olshen R, Cook EF, Sargent RK, Lamas GA, Dennis C, Wilson C, Deckelbaum L, Fineberg H, Stiratelli R.

A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain.
N Engl J Med. 1982 Sep 2;307(10):588-96

To determine whether data available to physicians in the emergency room can accurately identify which patients with acute chest pain are having myocardial infarctions, we analyzed 482 patients at one hospital. Using ***recursive partitioning analysis***, we constructed a decision protocol in the format of a simple flow chart to identify infarction on the basis of nine clinical factors. In prospective testing on 468 other patients at a second hospital, the protocol performed as well as the physicians. Moreover, an integration of the protocol with the physicians' judgments resulted in a classification system that preserved sensitivity for detecting infarctions, significantly improved the specificity (from 67 per cent to 77 per cent, P less than 0.01) and positive predictive value (from 34 per cent to 42 per cent, P = 0.016) of admission to an intensive-care area. The protocol identified a subgroup of 107 patients among whom only 5 per cent had infarctions and for whom admission to non-intensive-care areas might be appropriate. This decision protocol warrants further wide-scale prospective testing but is not ready for routine clinical use.

Discriminant Analysis

Blind use of stepwise [and 'same linear model forall'] in BOTH Discriminant Analysis & Logistic Regression !]

This technique (see ALR index) was more common early on, before logistic regression became widely available in packages.

The editorial "Statistical approaches to clinical predictions" McNeil BJ, Hanley JA. in the N Engl J Med. 1981 May 21;304(21):1292-4. commented on the use of discriminant analysis to identify patients with different types of gallstones . (one of these subgroups responds to longterm therapy, one does not)

Dolgin SM, Schwartz JS, Kressel HY, Soloway RD, Miller WT, Trotman BW, Soloway AS, Good LI. N Engl J Med. 1981 Apr 2;304(14):808-11.

Identification of patients with cholesterol or pigment gallstones by discriminant analysis of radiographic features.

In a search for a way to distinguish cholesterol gallstones from pigment gallstones by oral cholecystography, we evaluated 56 patients with surgically confirmed cholelithiasis. **Only buoyancy was highly predictive of gallstone composition: all 14 patients with floating stones had cholesterol stones** (P less than 0.01), but only one third of the patients with cholesterol stones had stone buoyancy. Using a function derived by stepwise discriminant analysis, we separated patients with cholesterol stones from those with pigment stones. The predictive accuracy was significantly improved: sensitivity was 95 per cent (37 of 39 patients with cholesterol stones), specificity was 82 per cent (14 of 17 patients with pigment stones), and efficiency was 91 per cent (51 of 56 total patients). The resultant function, applied prospectively to 17 additional cases, classified all of them correctly. In patients with cholelithiasis and gallbladders visualized on oral cholecystography, discriminant analysis can improve the prediction of gallstone composition and the subsequent selection of medial or surgical therapy.

Discriminant analysis is not that different in spirit from logistic regression (and indeed is the fore-runner of it). This particular dataset illustrated an interesting feature that shows the limitations of stepwise methods (discriminant analysis and logistic regression both): whereas **buoyancy was highly predictive of gallstone composition: all 14 patients with floating stones had cholesterol stones**, **it only entered 'second' in a stepwise analysis**. A clinician might well have separated the patients immediately on the basis of this first triage variable, then maybe even considered different other predictors for the 'floaters' and the 'sinkers'. In effect, this is like saying that another piece of information could have different usefulness in the two groups:- in the floaters it is not even needed, whereas if it does not, additional radiographic features can help refine the probabilities further.

This is why software for fitting these trees is sometimes called "AID" or Automatic Interaction Detection, rather than CART (Classification and Regression Trees)