Key Points in / Commentary on ALR

- **The multivariable case** ***(2.1 page 31, para 1)***

  I like this better than the often used term 'multivariate' (as in 'they controlled for V W X in a multivariate analysis'). In mathematical statistics, and in disciplines such as psychology, 'multivariate' means 'multiple Y values' e.g., a multi-dimensional responses variable based on say different scales, or multiple binary and other-scaled outcomes in a clinical trial (e.g., cure or not, side effect or not, good quality of life, ...) on the left side of the equation. Glantz in his regression textbook makes the same distinction between 'multivariate' and 'multi-variable'.

  So here, 'multivariable' means '2 or more X terms in regression'

  Incidentally, JH argues that one can learn most of the principles of 'multivariable' (i.e. multiple) regression from just 2 regressor terms.

- **Additional topic: design (indicator, dummy) variables for categorical independent variables** ***(2.1 page 31, para 1)***

  This brings up the distinction between a ***variable*** and ***the number of terms*** in the regression model needed to represent it. (as we will see, to represent a variable with k levels, we use k-1 indicator ***terms***. It explains why, if one had just one variable (e.g. 'race', with 3 categories), the regression is considered 'multivariable'. Race is 1 variable, but requires 2 indicators (plus the '1' associated with the intercept, to represent all 3 categories.

- **Coding schemes for categorical independent variables** ***(2.2 page 32)***

  None of this discussion is particular to logistic regression.. it is a 'right hand side of the equation' topic.

  By the way, JH's preference is that 'you make your own' indicator variables, rather than rely on the software to make them (where you cannot always get the reference value you want)

  Another issue at this stage: the coding scheme shown on p32 (having a reference category, and indicators for the other categories, is ***just one of many possible coding schemes*** . Watch out that the 'default' coding scheme used by your software is what you think it is (and the default can change from PROC to PROC in SAS, and is different in SPSS). The coding schemes are discussed in Chapter 3.

- **Fitting the multiple logistic regression model** ***(2.3 pp 33-36)***

  The relevant items to be highlighted in this section are:

  1. The idea of 'information' and the structure of the entries defined by equation 2.3 (on the diagonal) and 2.4 (off diagonal) elements of the 'information matrix'. The key is that the amount of 'information' about the parameters (a) is bigger if one has a bigger n (note that the information is a ***sum*** over the n observations) and (b) that information -- for ***logit*** regression -- is bigger if one is closer to π=0.5 and smaller if one is closer to π=0. To see this, examine the values of $\pi(1 - \pi)$ for various values π. At π=0.5, the product is 0.25, while at π=0.1, it is only 0.09. This is in contrast to the 'information' when dealing with/modeling π itself .. where the greatest variability and uncertainty (least information) is when one is near π=0.5, and greatest when near π=0 or π=1. In other words, ***the odds is most stable at π=0.5, where the odds is 1***. Sample proportions are least stable at π=0.5.
  2. think of information as the reciprocal of variance, and vice versa. Woolf's method of combining estimates from different strata was to use the reciprocal (inverse) of the variance, i.e. to use weights proportional to the amount of information.

     This explicit quantification of 'information' is a by-product of the Likelihood approach to estimation; the amount of 'information' in an observation is not so evident in other estimation schemes, such as least squares, where there is no probability model to help quantify things.
  3. Unfortunately, unless one has very well positioned values of the X 'vectors', the variance of a parameter estimate is not simply the reciprocal of the negative of the term in equation 2.3. This is because the estimation is for all parameters simultaneously, and the degree to which the different components of the X vector are collinear affects how precise the estimates are for the coefficient (parameter) associated with each component. So the inverse of the 'information matrix' is not readily computed by hand.
  4. The variance-covariance matrix of the beta_hats is obtained by adding a COVB option in the MODEL statement in SAS, or issuing the (post-estimation) 'vce' command in Stata. We already saw these for the simple logistic model in Chapter 1, since even though the equation had only 1 'X' variable (age in worked example, PSA in homework exercise), the regression equation had 2 terms (beta_0 and beta_1).
  5. If the x's in equation 2.3 were already 'centered' if each x had a mean of zero, then it is clear that the 'information' from an

observation with a large x (i.e. a value further from the mean) is more informative about the parameter than one near the middle. This is intuitive: to measure slopes more precisely, the most useful observations are those at the extremes of x.

- **Worked example: Low Birthweight** ***(2.2 pages 35 36)***

Before getting into details, take note of the overall proportion of low birthweight babies: 59/189 or 31% -- a VERY HIGH proportion. (usual in western populations is maybe 4-10%)

Notice also the 59:130 or 'odds' of 0.45:1. The log of this odds is –0.79 -- the starting point for our 'null' logistic regression, the same beta_0 coefficient obtained by simply fitting the model

In SAS	in Stata
PROC LOGISTIC DESCENDING;	logistic low
MODEL low = ;	logit

It is also helpful, at this point where the model is minimal, to calculate the logLikelihood, against which all improvements will be assessed.

If (without any covariates) everyone has the same 31% chance to be 'low', then the probability of observing the 59 : 130 split we did is

L = (0.31 to power of 59) x (0.69 to power of 130)

take log (natural) of this to get

logL = 59 x log(0.43) + 1`30 x log(.69) = –117.338

which, apart from rounding, agrees with the –117.336 reported by Stata or SAS.

- **Worked example: Fitted model ; Table 2.2** ***(2.3 page 36)***

We should spend some time inspecting (at least the signs of the coefficients in) Table 2.2 and try to draw the relationships implied by the fitted equation, both in the π scale and in the logit[π] scale

For example what if we plotted p_hat versus lwt ? against age ?
the fitted logits versus lwt ? against age ?

- **Testing for significance of the model** ***(2.4 page 36-40)***

A little precision of language would help here:

What the authors are referring to is an OVERALL (omnibus) test of

H_0: all 5 beta_s (i.e. all except beta_0) are ZERO

[effectively, that model is simply logit[π] = beta_0]

vs.

H_{alt}: one of more of 5 beta's is NON-ZERO

This is the same test that is paraded out as the overall F test in regular linear regression.. we are seldom interested in testing this H_0, and are often more interested in WHICH variables matter? HOW MUCH do they matter? HOW WELL do they triage the probabilities into those closer to 0, and those closer to 1 (as in diagnostic and prognostic functions) and [for confounding situations] WHAT is the adjusted coefficient of the 'exposure' of primary interest?

Yet, because of the way the test statistic is placed prominently on printouts, many 'hoping for something significant' researchers fix their gaze on the overall (global) test first, even though it is seldom their research focus.

- **Before concluding that (any or) <u>all</u> coefficients are non zero (middle of p*37)***

As you know from your other regression course, this is a tricky issue. Any such a statement must say which other terms are included in the model. And it can be misleading to look at the Wald test (each beta_hat divided by its SE) in a table such as Table 2.2 and conclude that 'lwt and possibly race were significant while age and ftv were not'

The individual test statistics and p-values in Table 2.2 are in relation to "variable entered last' hypotheses.. and the p-values could change radically if some of the other variables in Table 2.2 were not included in the model. And one shouldn't drop 2 variables at once: one should do things in a careful sequential order.. (if two variables were highly collinear, and the (common) variable they represent is an important predictor, one could obtain misleading (non-significant) p-values for each if both are included in the model.. and so dropping both, because neither is significant in a model that includes both, would be inappropriate.

The likelihood ratio test comparing the model with 6 terms (Table 2.2)

and 4 terms (Table 2.3) is a better way of testing the value of the two deleted variables age and ftv.

Note also the analogy here with partial F tests in regular regression and same "degrees of freedom for test = difference in size of models" idea

One compares the difference between the R-sq in a model with 6 terms, and its counterpart in a model with 4 terms using as a reference distribution, the F distribution with 2 degrees of freedom in the numerator -- and n-6 in the denominator.

In the limit, with a large n, under the null , 2 times the F[2, n-6] statistic has a distribution close to a chi-square distribution with 2 degrees of freedom -- i.e. the 2 tests are the same, With binary Y's, we don't spend degrees of freedom estimating a separate variance parameter for the variation of Y around its expectation-- we assume it is Bernoulli.. i.e. it is like having an infinite number of degrees of freedom in the denominator of the (partial) F statistic -- thereby making it like a chi-square statistic.

- **Likelihood Ratio test of a categorical variable** ***(2.4 page 38)***

This is an important warning.

Also, this is one situation where one cannot get by with the Wald test instead of the Likelihood Ratio test. For a <u>single</u> interval predictor variable, the results of the Wald test (which is a variable added last test) and the Likelihood Ratio test will often agree closely. Here, too, the Wald test is a z-test, and its square is therefore a chi-sq statistic with 1 df.
BUT since the (k-1) indicator terms for a (k-category) categorical variable are automatically correlated, one has to be doubly careful making decisions about any one category.. and one also has the 'arbitrariness of the coding scheme to contend with .. the Likelihood ratio test (of the null hypothesis that π is the same across all categories, all other variables being equal) is the same no matter what coding scheme is used (This is also a favourite phd exam question).

- **Wald and Score tests** ***(2.4 page 39)***

Say with Likelihood Ratio tests (& univariate Wald tests if applicable)

- **Confidence Intervals (still on logit scale)** ***(2.5 pp. 40-)***

You will need to be able to do such calculations by hand, in cases where your are interested in effect modification (interaction)

and so it is worth understanding the structure of eqn 2.7 and being able to do the calculations without getting muddled.. fortunately, the largest dimension you will have to deal with is usually 2, and not the 4 dimensional example in page 43.

What is involved here is the variance of a linear combination of random variables, using a rule that you probably had (in a simpler version) in 607

There, you probably saw

$$\mathrm{Var}[\, Y1 + Y2 \,] = \mathrm{Var}[Y1] + \mathrm{Var}[Y2] \ldots \text{ if } Y1 \text{ and } Y2 \text{ uncorrelated}$$

and you may have seen the more general version

$$\mathrm{Var}[\, Y1 + Y2 \,] = \mathrm{Var}[Y1] + \mathrm{Var}[Y2] + 2\,\mathrm{Cov}[Y1, Y2] \ldots \text{ general}$$

And, you may have seen the rule fro the variance of a weighted average of two uncorrelated r.v.'s

$$\mathrm{Var}[\, w_1.Y1 + w_2.Y2 \,] = w_1^2.\mathrm{Var}[Y1] + w_2^2.\mathrm{Var}[Y2] \ldots \quad \text{if } \mathrm{corr}[Y1,Y2]=0$$

or even the general case

$$\mathrm{Var}[\, w_1.Y1 + w_2.Y2 \,] = w_1^2.\mathrm{Var}[Y1] + w_2^2.\mathrm{Var}[Y2] + 2\, w_1 \,.\, w_2 \,.\mathrm{Cov}[Y1,y2]]$$

Then, equation 2.7 is just the general case of this, for a weighted sum of (p+1) random variables

beta_0_hat to beta_p_hat --

the estimated coefficients are the random variables,

and the x's used with them (in eqn 2.6) are the weights.

A useful way to keep the calculations straight is to form a 2 way table, for example with a linear combination of 3 random variables:

$$\text{Var}[\, w_1.Y1 + w_2.Y2 + w_3.Y3]$$

Remember that

Var[Y1] is just the covariance of Y1 with itself

Cov[Y1,Y2] is the same as Cov{Y2,Y1]

(Cov is a scaled correlation; correlations are symmetric).

Set the table up this way:

	w_1	w_2.	w_3
w_1	Var[Y1]	Cov[Y1,Y2]	Cov[Y1,Y3]
w_2.	Cov[Y1,Y2]	Var[Y2]	Cov[Y2,Y3]
w_3	Cov[Y1,Y3]	Cov[Y2,Y3]	Var[Y3]

Now , make the 9 products of the w in the row, the w in the column, and the variance or covariance entry in the covariance matrix (the part inside the border)

w_1^2 . Var[Y1]	w_1 . w_2 . Cov[Y1,Y2]	w_1 . w_3 . Cov[Y1,Y3]
w_1 . w_2 . Cov[Y1,Y2]	w_2^2 . Var[Y2]	w_2 . w_3 . Cov[Y2,Y3]
w_1 . w_3 . Cov[Y1,Y3]	w_2 . w_3 . Cov[Y2,Y3]	w_3^2 . Var[Y3]

Now , add the 9 products. You will see that the 3 diagonal elements are the first summation in eqn 2.7. You will also see that the entries above the diagonal have identical counterparts below it -- because of symmetry. So instead of adding 6 other products, you can add the 3 above the diagonal, and double their sum.

For the ugly worked example ion page 4, here is the scheme (if I haven't transcribed incorrectly .. it doesn't help that H&L write the equation with the beta_0_hat first but Table 2.4 has it as the last row and column):

	beta_0 1	lwt 150	ir_black 0	ir_other 0
1	0.7143	-0.005211	0.0226	0.1272
150		0.000041	-0.000647	0.000036
0			0.2382	0.0532
0				0.1272

It is hard to see on the page of arithmetic, but laying it out this way, with two of the weights being zero, means that we can dispense with all but the 4 products in the top left. They spared us the example of a 150 lb black woman, where we would have had

	1	150	1	0
1	0.7143	-0.005211	0.0226	0.1272
150		0.000041	-0.000647	0.000036
1			0.2382	0.0532
0				0.1272

and 9 products to keep track of.

AND, if you adopted another coding scheme for race, you cold have had 4 non-zero weighs, and 16 products to assemble and sum.

comments/corrections welcomed .. jh feb 1 2004