Key Points in / Commentary on ARL

- **Up from 2 x 2 table? table as special case of logistic regression?**
  (Preface page x, para 3)
  - even the 2 x 1 table (proportion + and – , no comparison, e.g. 3/20 in our first assignment) is a special case of regression .. one with just an intercept or 'overall proportion'
  - 2x 2 table (e.g. infections following warm surgery -- index category --vs. conventional surgery -- reference category) is a comparison of two proportions. and we have already compared them on SEVERAL scales,

    not just the proportion scale itself (Risk Difference),

    but also on the

    logit scale, where the difference in the log odds i.e. the difference in the logits, represents the log of the OddsRatio, and so the antilog or exponentiated difference in logits is the OddsRatio

    and on the

    log scale, where the difference of log risks is the log of RiskRatio.
  - 2 proportions [prevalences/risks] can be written as 1 master regression equation using as the 'regressor' variable (usually called X) an indicator of the index 'exposure' category (E = 0 if reference category and = 1 if index category)

    Risk[generic E value] = Risk[ref cat.] + RiskDiff if index cat.

    = Risk[ref cat.] + RiskDiff × E

    = B_0 + B_1 × E

    Check: Risk[E=0] = B_0 + B_1 × 0
    = B_0
    = Risk[ref cat.]

    Check: Risk[E=1] = B_0 + B_1 × 1
    = Risk[ref cat.] + Risk Diff

    = Risk[index cat.]

- **Up from logistic regression?**

  Logistic regression is just one example of Binary or Binomial regression. Bernoulli: each observation or line of data in file is for a ***single*** individual or trial, each response a 0 or 1 random variable; Binomial: the observation or line of data refers to a number of individuals (≥1) with the same covariate pattern, i.e. in the same "cell". The 'response' is the aggregated (0/1) results i.e. #pos/#persons in the persons with this profile or covariate pattern. Or, can think of Bernoulli = Binomial with just n=1 person in each "cell".

  Some Binary regression programs (e.g. PROC LOGISTIC in SAS) by default assume that the n for each observation (line of data) is 1 and so simply allow the user to specify MODEL Y = predictors without having to specify that 'n' = 1 (get an error if have a Y that is not 0 or 1, since if by default the denominator is 1, the numerator must be 0 or 1). PROC LOGISTIC also allows user to specify MODEL #pos/n = predictors , where you supply the numerator AND the denominator for the "cell"

  Other e.g. of Binomial regression is probit regression. SAS/Stata have a special program just for logistic regression (LOGISTIC / logistic) and another just for probit (PROBIT/probit).

  Q: Could all binary/binomial regressions be "rolled into one?"

  MORE GENERALLY Binary and Binomial regression are themselves special cases of Generalized Linear Models, where distribution can be one of SEVERAL (i.e. Gaussian, Bernoulli/Binomial, Poisson, Negative Binomial, Gamma, Inverse Gaussian) AND where different functions link the left side of the regression equation (the expected value [mean] of Y, at a given X, e.g. the expected proportion of Y=1 at this X value) to the predictors on the right side. Three examples are the IDENTITY link (e.g. link the proportions themselves to the determinant, as in risk difference, or link the LOG of the proportions, as for the RiskRatio, or the LOGIT, as for OddsRatio).

  In SAS and Stata these are called GENMOD and glm, respectively.

  This raises issue of whether we should learn a number of different programs, or just one general one (if the latter, we could do all the analyses in c621 and most of those in c681 from one generalized program, which simply "toggled" between the different distributions, AND the different shapes to the link between the left hand side (the expected value, or mean, or proportion) and the predictors

- danger of easy-to-use software (***(Preface page ix, para 2)***
  cf. story of statisticians &epidemiologists on train [link from FAQ]

Key Points in / Commentary on ARL Ch 1

- **Over last decade...standard method of analysis *(page 1, para 1)***

  (Given the mainly epi focus of the book) what is history of logistic regression? The c678 website has one of the first papers, in 1962 by the biostatistician responsible for 'the odds ratio in epi' [1], and logistic regression in epi. The stimulus for the 1962 paper was the Framingham study. (1) Cornfield J. A Method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst 1951;11:1269-1275.

- **Goal of model-building techniques in statistics.. find the best fitting and most parsimonious, yet biologically reasonable model to describe relation ship between (...) covariates. *(page 1, para 2)***

  Several purposes, such as

  -description (as suggested)
  -smoothing (for reliability)
  -ADJUSTMENT one 'covariate' is PRIMARY: the "exposure" (E);
  others are a nuisance, e.g. confounders,
  or create additional variation,
  or modify the relation between prob[outcome] & E

- **What distinguished logistic from linear.. Y is binary / dichotomous *(page 1, para 3)***

  As described above, in remarks on preface, there are a number of other regression models for binary outcomes .. they depend on the choice of link, or if you prefer, on the scale on which proportions are regressed on the determinants or covariates

  Careful re terminology as it relates to the "linear regression" model! "Linear" has several meanings. One relates to the makeup of ***right*** hand side (to be strictly technical about it, 'linear' refers to linear in the ***parameters***, ***not*** in the X's) . H&L's 'linear ' refers more to the ***left*** hand side, and the scale in which the regression is carried out (everything is same about the right hand side -- except maybe some constraints -- whether in 'logistic' or what they call 'linear' regression mode). The better way to distinguish models is via the LINK (the scale used in the left hand side) and the conditional DISTRIBUTION (Gaussian, et al for responses measured on a "continuous" scale, and Binomial, Poisson etc. for responses recorded on a binary or non-negative integer (count) scale.

- **Large Variability in CHD (Y) at t all ages (X). *(page 2, para 4)***

  1 This comes with the territory (the fact that we are measuring the response on a 2-point scale. This large uncertainty (entropy) is what makes the outcome of baseball, football, hockey interesting, but frustrating for people concerned with other individual (present of future) states/outcomes (e.g. physicians, weather forecasters, admissions committees, etc.).

  One should however, distinguish between predicting the mean response or proportion of responders in the (sub)universe of persons with a particular covariate pattern, and predicting the response of the individual.

  For example, if a weather forecaster says the probability of a certain type of weather is 60% and on the (one) day in question, that type of weather shows up, is this a good prediction? What if it doesn't turn out that way that day? Or is the forecaster a good prognosticator if in 60% of the days in which he says the probability is 60%, and in 30% of those days in which he says the probability is 30% that type of weather does show up, etc.. ?

  2 The technique of 'creating intervals for the independent variable' to 'remove some variability' is a form of smoothing, or 'borrowing strength' from neighbouring individuals. Moving averages (where one slides the interval, usually a ***time*** window), are a variation on this 'smoothing' idea.

- **It appears that as age increases, the proportion of individuals with evidence of CHD increases. *(page 2, para 5)***

  This is a 'pet peeve' ('a particular source of aggravation') of mine. These data come from a cross-sectional study, and so the proportions represent prevalences. The statistically correct way to describe this prevalence pattern is "it appears that the prevalence of CHD is higher in older individuals". If I had a dollar for every time I heard this 'dynamic' interpretation of a cross-sectional relationship, I would be rich. I think this idea of 'as X increases, so does Y' comes from the same 'physics lab' mentality that makes us call X the 'independent' variable and Y the 'dependent' one. This might be the case in the lab, where one freely turn knobs or dials to manipulate temperature, lighting, humidity, etc.. and then observe the response of a variable that truly 'depends' on or 'responds' to' these changes in a causal way. Unfortunately, most epidemiology data is not derived from such experiments.. Nor can we freely (and independently?) manipulate the knobs and dials in any combination we wish.

- **In any regression problem, the key quantity is the mean value of the outcome variable, given the (i.e. as a function of) the independent variable** ***(page 4, para 1)***

  d'accord. I agree 100%. But, to emphasize this ***even more***, I would take out the 'E' in the notation E{Y | x] and replace it by a μ, i.e.

  the pattern of μ[Y | x] 's as a function of x

  After all, when ones eye, on looking at a scatter plot, mentally draw a line or curve through the data, one is forming a line (or curve) of middles". Indeed, one can then take away the data (noise) and be just keep the (estimated) signal.

- **With dichotomous data, conditional mean must be between 0 and 1;** ***(page 5, para 1)***

  Here, why not change directly to proportions, and use the parameter π, for proportion, or if you are less pretentious, the upper case P ?

  Now the challenge is to describe

  the pattern of π[Y=1 | x] 's as a function of x

  In our own exercises in assignment 1, we ran into several situations and limits where we were dealing with a proportion near 0 or 1, and in some of these we switched to a logit scale (remember the first class, where we had the probability scale as a kind of thermometer, and beside it, a number of other scales, including the open-ended logit (log odds) scale, with extremes of minus and plus infinity.

- **Many (cumulative, or 'distribution)' functions have been proposed** ***(page 6, para 1)***

  A popular S-shaped cumulative curve, especially before the logistic, and still popular with toxicologists, was the probit curve.

  The reasoning in toxicology was that the tolerances of (the smallest amount of poison that would kill) individual animals has a Gaussian distribution (on some concentration scale). Then if one gave a dose that was 2 SD below the average on this scale, one would kill 2.5%; if one gave a dose that was 1 SD below, one would kill approx. 16%; if an average dose 50%; a dose 1 SD above would kill approx. 84%, and 2 SD above would kill 97.5%. And so on.

- **Two primary reasons for choosing logistic distribution (curve)** ***(page 6, para 1)***

  First: (as indicated) The probit used to be more difficult to work with. this is lees the case nowadays, now that we have generalized linear models.

  [aside] Some statisticians use the logistic distribution as a way to draw close-to-Gaussian random variables for Monte-Carlo simulations (unlike the logistic, the Gaussian c.d.f. function does not have a closed form).

  Flexibility (of shape): for a good part of the (0,1) range, the logit and probit are quite close to each other. They differ more at the extremes of the (0,1) scale.

  Second [interpretation] This is particularly pertinent in epidemiology, where, in of case-control studies, we are 'stuck with' the odds ratio.

  The logistic form also arises naturally in discriminant analysis (indeed, it was from discriminant analysis that Cornfield derived, and others refined, the famous Framingham logistic) risk function.

- **Equation 1.1 (the logistic regression model)** ***(page 6)***

  This is the form if you work in the (0,1) scale

  the equation further down is the one that looks more like a regression equation i.e.

  some function of Prob[Y=1 | x] is a lin. combination of x's and β's

  I would have written it with the ln[p/(1-p)] leftmost, to emphasize that one is altering the scale on the left hand side.

  ' logit transform of ' $\pi[Y=1 \mid x] = \beta_0 + \beta_1 x$

  Notice the ln for 'the natural' log , i.e., log to base e.

- **2nd important diff: distribution of Y conditional on π[Y=1 | x]** ***(page 6/7)***

  This is a binomial with an 'n' of 1 (i.e., a Bernoulli random variable]

- **Maximum likelihood Fitting of logistic regression model)**
  (section 1.2 page 8)

In the method of Least Squares, there was no intrinsic assumption about the distribution of the errors (the ε's, i.e. the variations of the individual Y's at a given value of X from the true mean of all possible Y's at this same given value of X) . i.e. the least squares criterion is a strictly mathematical or geometrical criterion for closeness of the fit of the data points to the line. the fact the method of least squares has some good properties if in fact the data follow certain laws is another matter)

By contrast, the method of Maximum Likelihood requires a specific statistical model for the distribution of the observations. It is not enough that the model specify the mean value of Y at each X, but also the exact pattern of variation of these x-specific Y's about that mean.

Two examples might help explain the entirely different approach implied by the use of the maximum likelihood criterion. in the first of these, we consider just two values of the parameter of interest, in the other we consider all possible (positive) values of a regression slope

1. cf. the spreadsheet on the colour distribution of a sample of M&M candies from one of two sources (data and relevant formulae are in sheet 1 of the Maximum Likelihood spreadsheet). They are all from one or other of two sources: either 'milk chocolate', where 70% of the source have certain colours; or 'crispy chocolate', where 50% are of these colours. We don't know which, and we have no prior information to guide us one way or the other (we are neutral), and we cannot use our sense of smell, just our sense of sight.

   the observed proportion of the colours in question is 60%. So, are the data more likely to have come from the source with 70% or the source with 50%

   the question effectively asks you to calculate the probability of observing 60% under each of the 2 scenarios (the two competing parameter values are 'milk' i.e. 70&, and 'crispy' i.e. 50%).

   and to rank the two sources by this 'probability of obtaining the observed data, or what is called the "likelihood".

   Moral: even though 60% (data) is equidistant from 50% and 70%, the data are more likely under the '50% in source' model. (the 50% model also gives the smaller chi-squared value too, if we adopt a minimum-chi-square criterion.

2. cf. the spreadsheet on the numbers of errors on two manuscripts of 1 and 2 pages respectively (data and relevant formulae are in sheet 2 of the Maximum Likelihood spreadsheet). The simple regression model is very 'simple', it does not even have an intercept. the model is simply

   $E[\ \#\ \text{errors} \mid \#\ \text{of pages}\ ] = \beta * (\#.\ \text{of pages})$

   i.e.

   $E[\ Y \mid X\ ] = \beta * X$

   what is your regression estimate of β if you estimate it by

   least squares ? (say proc reg, with NOINT option)

   grade 4 math?

   Maximum Likelihood [need complete statistical model] ?

The different methods imply different metrics..

Least squares measures discrepancies between a y and its (fitted) expected value (line) by the square of the (y – expected), and the overall closeness as the sum of these squares, summed across the datapoints.

The method of Maximum Likelihood measures the closeness of an observed y to a proposed expected (fitted) value using the height of the theoretical probability distribution function at that y value.. and the overall closeness of the datapoints by the product of these probabilities. For any (trial) value of β, one determines the implied fitted values. These then become the means for the probability model. In this case one might be tempted to adopt as a model the Poisson distribution, so that for each manuscript size X, one has with the fitted (proposed) mean $\mu = \beta * X$. Say we suggest as an estimate the value $\beta = 2.1$, so that the y=2 and y=5 are then calculated as coming from means of 2.1 and 4.2. The probability of obtaining a 2 in a Poisson distribution with mean = 2.1 is 0.27; the probability of obtaining a y=5 from a Poisson distribution with a mean of 4.2 is 0.163. The probability of obtaining the 2 and the 5, when β=2.1, is therefore 0.27 x 0.163 = 0.044. One would now change to a new candidate value of β and recalculate the probability of the observed 2 and the 5 under the two new means implied by this new value of the parameter β. Thus, one can obtain a curve of

the probability values as a function of β, a function that is called the Likelihood, one can obtain that β value which makes the probability largest. This is called the Maximum Likelihood Estimate (MLE) of β, or the value of β that makes the data more likely than any other value of β.

Of course, one might well be able to find the MLE by more direct and less 'brute force' methods. In some instances, as in this very one, one can take the derivative of the Likelihood function with respect to β and find the value of β where this is zero directly, by algebra Or, if need be, one can use a computer search for where the Likelihood is a Maximum.

One may have observed that when one multiplies a large number of probabilities together, their product becomes small, and so we more often work with the log of the product, i.e. the sum of the logs of the probabilities for the individual data points. This has two advantages: it shows the fact that one is summing over (aggregating information from) the different observations, and it turns out that the log of many probabilities derived from the standard distributions is easier to work with than the probability itself (a case in point is the height of the Gaussian curve at the observed value y. The height is

proportional to exp[minus the square of (y - fitted y)],

so that the log of the probability (or likelihood contribution) is

proportional to minus of square of (y - fitted y)

and indeed this shows why the general (maximum likelihood) method of estimation that H&L refer to on para 1 of page 8 does in fact coincide with the method of Least Squares when the error terms have a Gaussian distribution -- the likelihood will be at a maximum when the negative of the sums of the squares of the (y - fitted y) 's is at a maximum, i.e. when the positive of the sums of the squares of the (y - fitted y) 's is at a minimum ('least' squares).

This maximization, by working in the log Likelihood scale, and setting its derivatives to zero, is what is summarized in equations 1.4 to 1.6, for the case of a Binomial model. In their example, unlike our simpler zero-intercept model, there are two parameters β_0 and β_1, (and the sometimes write them as a single bold β, a vector) so that the search is a 2-dimensional search, just as if a blind person were finding the highest point on Mont Royal by repeatedly using a cane, and the gradients in all directions around him/her, to decide in which direction led uphill. Then by iteration, one would find the summit (provided one did not go up the second-highest peak of Mont Royal by mistake). Fortunately, in most Likelihoods, the log Likelihood is like an upside down wok (the mathematicians call it concave upwards) i.e. it does not have any dents or valleys, and cannot fool the summit-seeker by a 'local' maximum.

- **An interesting consequence: Σ y = $\Sigma \hat{\pi}$**
  (page 10, 1st para)

  This is the same (Sum of observed frequencies = Sum of expected, or fitted frequencies) constraint on the frequencies in a frequency table that takes away some degrees of freedom when testing the fit of the observed and expected frequencies.

- **Worked example, data in Table 1, Fitted parameters in Table 1.3.**
  (page 10)

  See under resources for the SAS/Stata programs that reproduce the results in Table 1.3, and that plot the observed points shown in Table 1.2, along with the fitted regression equation 1.7.

  Overleaf is a contour plot of the log Likelihood, confirming the MLE's of -5.3 for the intercept (β_0) , and +0.11 for the slope (β_1)

  Note also a second version, with age 'centered' at 45, and how much easier the search is (the parameter estimates are now not so negatively correlated, since a perturbation in the slope does not have so much influence on the intercept.

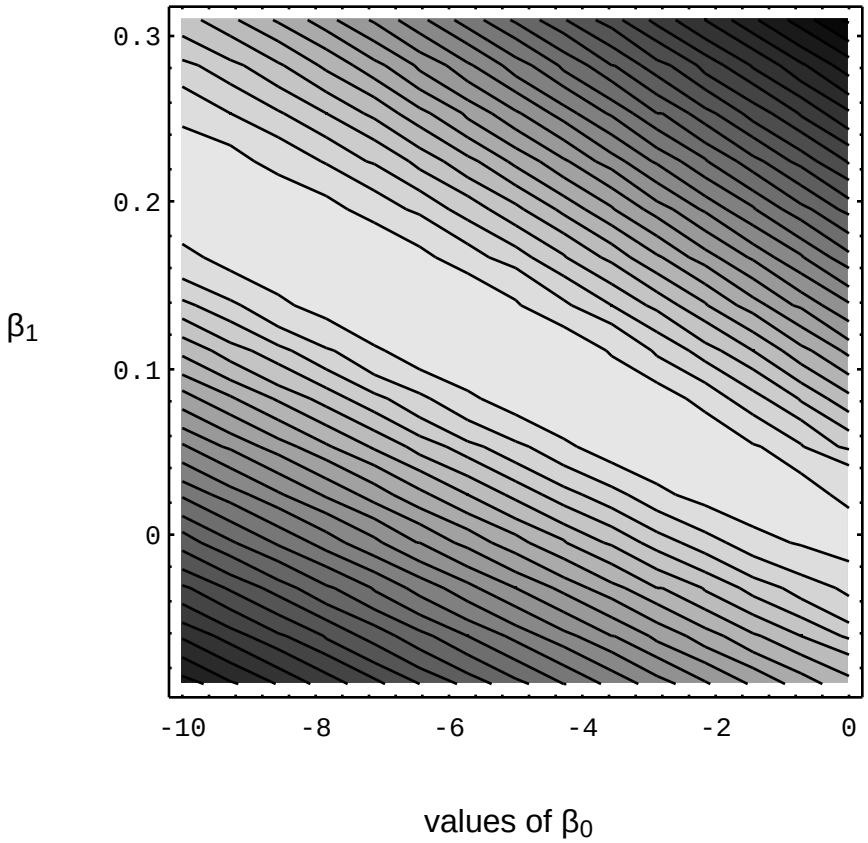
- **Testing for the significance of the coefficients**
  (page 11- section 1.3,)

  Note the operationalization of this, in the question posed in italics in the middle of page 11. Note also that the authors emphasize that they are not, at this stage, asking about the 'fit' in an absolute sense, but rather in a ***relative*** sense.

  At bottom of page 12, again in italics, is the criterion used in Likelihood-based fits. It might have been clearer if they had said

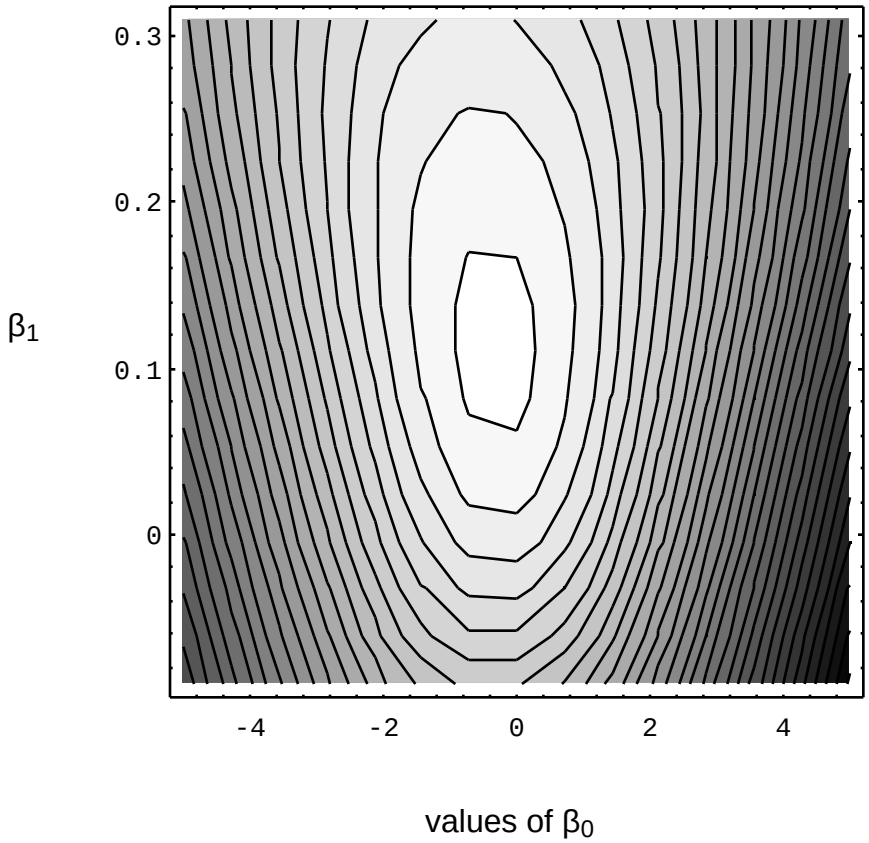
  Compare the product of the n ***probabilities*** of the observed values under the two models -- i.e. with and without the extra predictor variable. In practice, compare the sum of the logs of the n individual probabilities i.e., the log-likelihood.

Contour plot of the **log Likelihood**, a function β_0 and β_1.



Contour plot of the **log Likelihood**, a function β_0 and β_1,

but with age 'centered' at 45. Note that the MLE of the intercept changes, but not that of the slope, and that the parameter estimates are now much less correlated



- **"Saturated" model in this example..** ***(page 13)***

  In this example, the saturated model is one where you can perfectly predict for each person whether they had CHD or not. It is like having another column of data, such as the results of the actual tests (or self-reports, or whatever was used in this dataset) that classified the person as having CHD

  Or think of it as being able to have a probability 'curve' for Figure 1.1 that bounces up and down as needed and goes through every observed 0 or 1!

- **NB Deviance reported by SAS and by Stata** ***(page 13)***

  To e what they are talking about, compare your output with that obtained by others using a different package (or even GENMOD vs. LOGISTIC)

  Ultimately though, since , as H&L say, –2 times the log likelihood of the saturated model cancels out when you subtract the deviances of two realistic models of different sizes, this discrepancy between packages is not serious in practice.

  Also, at a practical level, the difference in deviances for two models is usually not identified in computer printouts as G (see 1.12 p 14). G seems to be mainly a textbook notation. In printouts, you may see –2logLR, i.e. minus twice the log of the Likelihood Ratio

  Some packages report the –2logL and let you subtract them yourself for two different models. Some report –logL, so watch out

  You can see why G is a ***difference***.. the log of a ***Ratio*** is the ***difference*** of the components.

  Also, how do you know if a log likelihood is better if small or if large?

  Think back to the likelihood of a saturated model. In this example, the probability of getting the observed Y would be 1, and so their product (over all n observed Y's) would be 1. Therefore the log Likelihood would be 0. An imperfect model would have a product that was less than 1, and so a log likelihood that is negative.. The more negative the log likelihood, the worse the model. But if you work with the ***negative*** log Likelihood, the ***larger*** the –logL (the more it is above 0), the worse

  Remember also that , just like a SSR, or an SSTotal, its size of logL is a function of the number of observations.. L gets smaller the more probabilities you multiply. So logL is a greater negative value if based on more data. So, be careful if you compare two logL's (or 2logL's) from different models, if the numbers of observations used to calculate the two are not the same (This can happen if with the larger model, some observations are dropped because they are missing data on the additional variables in the larger model).

- **"Wald" test..** ***(page 16)***

  In large datasets, with a lot of information (large numbers of subjects with Y = 1, large numbers with Y=0, and not too much collinearity in the X's) this difference between the Wald test and the Likelihood Ratio test is usually not that great.

  Or think of it as being able to have a probability 'curve' for Figure 1.1 that bounces up and down as needed and goes through every observed 0 or 1!

- **Confidence Interval Estimation** ***(page 17- section 1.4,)***

  The large CI for the 'far away' intercept estimate is no surprise.. if your age data are a long ways from 0, you can't expect that projections from where the data are, back to 0, will be precise.

  This is particularly the case when the 'X" is calendar time, using AD's such as 1998, 2000, 2003, etc.. Then projecting back to 0 Ad is quick imprecise, as well as being silly.

  The contours earlier show the better way to estimate things..

  They also show that if you want to minimize the covariance in equation 1.18, you would be well advised to centre your data near the X values of greatest interest.

  The calculations on pp 19 and 20 are very like what we have done already when calculating CI's for Risk and Odds Ratios -- calculate the CI in the log or logit scale, THEN convert the 2 limits of the CI back to the limits in the desired scale.

- **Other estimation methods** ***(page 21 - section 1.5,)***

  One interesting approach, used by Berkson ('of the bias'), was a minimum chi-square approach, used a lot (as was probit analysis) for fitting toxicology curves, with several observations at each dose (like age-grouped example in Table 1.2). H&L use the same chi-sq to compare observed and fitted values, but with MLE fits .

  The discriminant analysis approach was overtaken by logistic soon after 1980. see editorial about it versus logistic regression to identify which gallstones do / do not respond (only know after about 6 months!) to medical treatment. McNeil BJ, and JH: "Statistical approaches to clinical predictions". NEJM 304: 1292-1294, 1981.