11. Analysis of Case-control Studies
Logistic Regression

This chapter builds upon and further develops the concepts and strategies described in Ch.6 of Mother and Child Health: Research methods.

We have so far considered situations where the outcome variable is numeric continuous and the explanatory variables are a mix of numeric and categorical ones. But in clinical work we often wish to evaluate the effects of multiple explanatory variables on a binary outcome variable. For example, the effects of a number of risk factors on the development or otherwise of a disease. A patient may be cured of a disease or not; a risk factor may be present or not; a person may be immunised or not, and so on. The situation now is different from multiple regression analysis which requires that the outcome variable be measured on a continuous scale. For example, in the data file on Crime in the Unites States, the outcome variable 'Crime Rate' was continuous, and among the explanatory variables one (North/South) was categorical, the rest being all numerical. In the data file on Scores, the outcome variable 'Score' was continuous, and both the independent variables (Sex and Teaching Method) were categorical. When the outcome variable is binary, and one wishes to measure the effects of several independent variables on it the method of analysis to use is Logistic Regression. The binary outcome variable is coded 0 and1. The convention is to associate 1 with 'success' (e.g. patient survived; risk factor is present; correct answer is given, and so on), and 0 with 'failure'.

For Logistic Regression the following conditions need to be satisfied:

1. An outcome variable with the two possible categorical outcomes, preferably in the form 0 and 1. (A dummy variable may have to be created for the purpose).
2. We need to estimate the probability P of the observed value of the outcome variable out of 1.
3. We need a way of relating the outcome variable to the explanatory variables. This is done by means of the logistic function, which is explained later.
4. We need a way of estimating the coefficients of the regression equation developed.
5. We need to test the goodness of fit of the regression model, as well as estimate the confidence intervals of the coefficients.

Recall that in multiple linear regression the basic activity is to draw the least square line around which the values of Y (the outcome variable) are distributed. In contrast to that, in logistic regression we are trying to estimate the *probability* that a given individual will fall into one outcome group or the other. Thinking in terms of probability helps to interpret the coefficients in the logistic regression model in a meaningful manner, in the same way as we attach meaning to the coefficients in linear regression.

As a measure of the probability we ask about the odds of an event occurring. If P is the probability of a given event, then (1 – P) is the probability of the event not occurring. The odds of the event can be defined as:

Odds = $p/1-p$ i.e. Probability of event occurring ÷ Probability of event not occurring.

In other words, P / (1 – P) is the ratio of the probability of success, also called the *odds* of success. (When given the odds ratio, we can calculate probability from P = odds ratio ÷ (1 + odds ratio).

If there are several factors involved in deciding the eventual outcome we can calculate the odds ratio for each one separately. (This kind of crude odds ratio was described in relation to case - control studies in Chapter 6 of Mother and Child Health: Research methods). The joint effect of all the independent variables put together may be expressed mathematically as:

$$\text{Odds} = P / (1 - P) = e^{\alpha + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta pXp}$$

Where *e* is a mathematical expression (like π). The numerical value of *e* is equal to 2.71828.

The independent contribution of each of several factors (i.e. X_1, X_2 … X_p) in the above expression to the overall odds ratio can be determined by taking the logarithms of both sides of the equation.

$$\text{Log}_e\{P / (1-P)\} = \log e^{\alpha + \beta 1X1 + \beta 2X2 + \ldots \beta pXp}$$

The term log {P/(1–P)} is referred to as 'logistic transformation of the probability P, and is written as logit (p), which is short for 'logistic unit'. Transforming the counted proportion p (so many successes out of so many trials) to logit gets rid of the drawback of probability which varies from 0 to 1, whereas the logit can vary from –∝ to + ∝. And so we have now natural logarithm of (P / 1– P) = logit P = $\alpha + \beta_1X_1 + \beta_2X_2 + \ldots\ldots\beta_pX_p$

The right hand side of the equation is the same as in multiple linear regression with which we are familiar. Recall that in linear regression the parameters of the model are estimated using the *least squares* approach. What happens is that the coefficients selected are such that the sums of squared distances between the observed values and the predicted values (i.e. the regression line) are smallest.

In logistic regression the parameters are estimated such that the coefficients make our observed results most likely. The method is called the *maximum likelihood* method. What does it all mean? Suppose in the above equation the two possible values of the outcome variable are presence or absence of disease, (with presence of disease =1). Then α represents the overall disease risk, while β_1 is the fraction by which the risk is increased (or decreased) by every unit change in X_1; β_2 is the fraction by which the disease risk is altered by a unit change in X_2, and so on. The independent variables can be quantitative or qualitative e.g. by using dummy variables. However, one should bear in mind that the logit of a proportion p is the logarithm of the corresponding odds. If an X variable has a coefficient β, then a unit increase in X increases the log odds by an amount equal to β. This means that the odds themselves are increased by a factor of e^{β}. So if β = 1.6, the odds are $e^{1.6}$ = 4.95.The logarithms of all the odds can be converted back to probability by the formula P = $e^y/(1+e^y)$.

With this introduction about the principles of logistic regression we next examine how it works in practice by considering the following example.

Several tests have been developed for diagnosing coronary artery disease. Cardiac catheterization is the most definitive of these tests, but also most invasive. A non-invasive method is imaging the blood flow to the myocardium using the radioisotope thallium while the patient is made to exercise.

100 people (55 men and 45 women) had both the exercise thallium test and cardiac catheterisation. Some of them were on propranolol. Change in heart rate and E.C.G. during exercise were monitored, and occurrence of pain during exercise was recorded. The objective was to determine which variables were good predictors of a positive thallium test.

The data are given below.

In the data set the following codes apply:

Thallium (1= positive scan; 0 = negative scan) is the outcome variable.

The explanatory variables are:
Degree of stenosis of coronary artery as percentage block (0% = no block; 100% = complete block)
Use of propranolol prior to test (0 = no; 1 = yes).
Maximum heart rate during exercise (0 = heart rate did not rise to 85% of maximum predicted rate; 1= heart rate exceeded 85% of predicted rate).
Ischaemia during exercise (1 = occurred; 0 = no).
Sex (0=male; 1=female).
Chest pain during exercise (0 = no pain; 1 = moderate pain; 2 = severe pain).
E.C.G. changes during exercise (0 = no; 1 = yes).

(*Am.J.Cardiol.1985;55:54-57*)

Thall	Stn	Propran	HrtRate	IscExr	Sex	PnExr	ID
1	40	0	0	0	1	0	1
1	40	0	1	0	1	0	2
1	40	1	0	1	1	2	3
1	30	0	1	0	1	0	4
1	30	0	0	0	1	0	5
0	40	0	1	0	1	1	6
0	30	1	1	1	0	0	7
0	30	0	1	0	1	2	8
1	40	1	0	0	0	0	9
1	30	0	0	0	1	0	10
1	30	0	0	0	0	0	11
0	30	0	0	0	1	0	12
1	40	0	0	0	0	0	13
0	0	0	1	0	0	1	14
1	0	0	1	0	0	2	15
0	0	1	1	1	0	0	16
1	0	1	1	0	0	0	17
0	0	0	1	0	1	0	18
0	10	0	1	1	0	0	19
0	0	0	0	0	1	0	20
0	0	0	1	0	0	1	21
1	0	1	0	0	1	0	22
0	0	0	1	0	1	1	23
0	0	0	1	0	0	0	24
0	0	0	1	1	1	0	25
0	0	0	1	0	1	0	26
0	0	0	1	1	1	0	27
0	0	0	1	0	0	0	28
0	0	1	1	1	1	0	29
0	0	0	0	0	1	0	30
0	0	0	1	0	1	0	31
0	0	1	0	1	0	0	32
0	10	1	0	0	0	0	33
0	0	1	1	0	1	0	34
0	0	1	0	0	1	0	35
1	0	0	0	1	1	0	36
0	0	0	1	0	1	0	37
0	0	0	1	0	1	0	38
1	0	1	0	0	0	2	39
0	20	1	0	1	0	0	40
0	0	1	0	1	0	0	41
0	0	1	0	1	0	0	42
0	0	1	1	0	0	0	43
0	0	1	1	0	1	0	44
0	0	0	1	0	0	0	45
0	0	1	1	0	1	0	46
0	0	1	0	0	1	0	47
0	0	0	1	0	0	0	48
1	0	1	0	0	0	0	49
1	0	1	1	0	0	0	50
0	0	0	0	0	1	0	51
1	0	0	0	0	0	0	52
0	0	0	1	0	1	0	53
0	0	0	1	0	1	1	54
1	0	1	1	0	1	2	55
1	0	1	0	0	0	0	56
0	0	0	1	1	0	0	57
0	0	0	0	0	0	0	58
0	40	1	0	0	0	2	59
0	0	1	0	0	0	0	60
0	0	1	0	0	0	0	61
1	0	1	1	1	0	1	62
0	0	1	0	0	0	0	63
1	30	1	1	0	1	0	64
0	20	1	0	0	1	1	65
0	30	1	0	0	0	0	66
1	0	1	0	0	0	0	67
1	30	1	0	0	1	0	68
0	0	1	1	0	1	0	69
1	0	1	0	0	0	0	70
1	0	1	1	0	1	0	71
1	0	1	1	0	1	0	72
Thall	Stn	Propran	HrtRate	IscExr	Sex	PnExr	ID
1	0	1	0	0	1	0	73

0	30	1	1	1	0	0	74
0	50	1	1	0	1	0	75
0	0	1	0	0	0	0	76
1	0	1	0	1	0	1	77
0	0	1	0	0	0	0	78
0	0	0	0	0	1	0	79
0	0	0	1	0	0	0	80
0	0	0	0	0	1	0	81
0	0	0	1	0	0	0	82
1	0	0	0	0	1	1	83
0	0	0	1	0	0	0	84
0	0	0	1	0	1	0	85
0	0	0	0	0	1	0	86
0	0	0	0	0	1	0	87
1	0	0	1	0	0	0	88
0	0	0	1	1	1	0	89
0	0	0	0	0	1	0	90
0	0	0	0	0	1	0	91
0	0	0	0	0	1	0	92
0	0	0	1	0	1	0	93
0	40	1	0	0	1	0	94
1	40	1	1	1	1	0	95
1	40	0	0	0	0	0	96
0	0	0	1	0	1	0	97
1	0	0	1	0	0	2	98
1	0	0	1	0	1	0	99
0	10	0	0	0	1	0	100

We next perform the logistic regression analysis with Thallium as the binary outcome variable. Recall that Thall = 1 means positive scan; 0 = negative scan.

Binary Logistic Regression

[In MINITAB Stat → Regression → Binary Logistic Regression]

```
Link Function:  Logit

1).Response Information

Variable  Value       Count
Thall     1              34  (Event)
          0              66
          Total         100

2).Logistic Regression Table
                                                        Odds        95% CI
Predictor       Coef      StDev        Z      P      Ratio    Lower    Upper
Constant     -0.9349     0.5165    -1.81  0.070
Stn          0.03080    0.01482     2.08  0.038       1.03     1.00     1.06
Propran       0.6000     0.4844     1.24  0.215       1.82     0.71     4.71
HrtRate      -0.4234     0.4735    -0.89  0.371       0.65     0.26     1.66
IscExr       -0.6322     0.6601    -0.96  0.338       0.53     0.15     1.94
Sex          -0.2996     0.4780    -0.63  0.531       0.74     0.29     1.89
PnExr         0.6953     0.4009     1.73  0.083       2.00     0.91     4.40

3).Log-Likelihood = -57.650
Test that all slopes are zero: G = 12.907, DF = 6, P-Value = 0.045

4).Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   60.350    38  0.012
Deviance                  66.811    38  0.003
Hosmer-Lemeshow           14.243     6  0.027

5).Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)
```

```
                                Group
Value      1     2     3     4     5     6     7     8   Total
1
  Obs      2     1     2     5     5     6     9     4      34
  Exp      2.4   4.1   2.9   3.1   3.7   5.0   5.7   7.0
0
  Obs     15    18     9     5     5     6     2     6      66
  Exp     14.6  14.9   8.1   6.9   6.3   7.0   5.3   3.0

  Total   17    19    11    10    10    12    11    10     100

6).Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs           Number  Percent     Summary Measures
Concordant        1664    74.2%     Somers' D                0.51
Discordant         524    23.4%     Goodman-Kruskal Gamma    0.52
Ties                56     2.5%     Kendall's Tau-a          0.23
Total             2244   100.0%
```

A number of diagnostic graphs are available for checking the appropriateness of the model. These are described in the section "Logistic Regression Diagnostics".

Interpreting the results

The different parts of the output have been numbered for ease of following the explanation.

1). **Response Information** provides information about the response variable Thallium. There were 34 subjects with a positive scan (Value = 1), described here as event.

2). **Logistic regression table** shows the estimates of the β coefficients their standard deviations, z values (z = β coefficient ÷ standard deviation; also called Wald Statistic), and p values. Also the odds ratio and 95% confidence intervals for each of the coefficients are printed.

In order to understand the logistic coefficients let us first consider what is being tested. We are looking at the odds of events. By definition, the odds of an event occurring is the ratio of the probability that it will occur to the probability that it will not occur. In other words Probability (event) ÷ Probability (no event).The logistic model is about the logarithms of the odds, which is referred to as logit. So the model may be written as:

$$\text{Log}_e\{\text{Probability (event) / Probability (no event)}\} = \text{Log}\{e^{\alpha + \beta 1X1 + \beta 2X2 \ldots \beta kXk}\}$$

$$= \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_k X_k$$

From the equation we can interpret that the logistic coefficients are measures of the change in log odds (i.e. logit) associated with a one unit change in the explanatory variable if it is a continuous numerical variable. Thus the coefficient of Stnprcnt (% stenosis) gives the estimated change in
log { P(+ve thallium scan) ÷ P(–ve thallium scan)} for every 1% stenosis as determined by coronary angiography whilst holding other independent variables constant. And $e^{\beta_{value}}$ gives the actual odds ratio. For example, $e^{0.03080} = 1.031279$; $e^{0.6} = 1.822$, … and so on.

Significant *P* value at the bottom of the regression table tells us that there is a significant association between at least one explanatory variable and the outcome by testing whether all slopes are equal to zero. If the *P* value were not significant there would be no need to go

further. Next we consider the *P* values for each term in the model. These values tell us whether or not there is statistically significant association between a particular explanatory variable and the outcome.

3). **Log Likelihood** tells us about the probability of the observed results given the β coefficients. Since the likelihood is always less than 1 it is customary to use the logarithm of the likelihood and multiply it by –2. Hence the term –2LL. Its value provides a measure of how well the model fits the data. The log likelihood statistic is analogous to the error sum of squares in multiple linear regression. As such it is an indicator of how much unexplained information remains after fitting the model. A large value means a poorly fitting model. The larger the value of the log likelihood the more unexplained observations there are. Therefore, a good model means a small value for –2LL. If a model fits perfectly, the likelihood is 1, and $-2 \times \log 1=0$. One advantage of multiplying the log likelihood by 2 is that -2LL has an approximately chi square distribution and so makes it possible to compare values against those that arise by chance alone.

Next to consider is the statistic G. G is the difference in -2 LL between a model that has only the constant term and the fitted model. G tests the null hypothesis that all coefficients of the independent variables equal zero (i.e. no slope) versus the coefficients not being equal to zero. In our case the *P* value is 0.045, showing that at least one of the coefficients is not equal to zero.

4. **Goodness-of-Fit**. How well the model fits the observed data is assessed by a number of ways as follows:

- A variety of statistical tests (Pearson; Deviance; Hosmer-Lemenshow) are applied with their *P* values determined in order to assess how well the model describes the data. A low *P* value indicates that the predicted probabilities deviate from the observed probabilities in a way that the binary distribution does not predict. For example, the Hosmer-Lemenshaw test assesses the fit of the model by comparing the observed and expected frequencies. The estimated probabilities are grouped from lowest to highest, and then the Chi Square test is carried out to determine if the observed and expected frequencies are significantly different. When the Hosmer- Lemenshaw test is significant (as it is here) it means that the observed counts and those predicted by the model are not close, and the model does not describe the data well. When the Hosmer-Lemenshaw test is *not significant* it means that the observed and the predicted counts are close and the model describes the data well. In our case here the *P* values range from 0.003 to 0.027. This means that the null hypothesis of data not fitting the model cannot be rejected. In other words the goodness-of-fit is not that good.

  Pearson and Deviance are both types of residuals. The larger the *P* value the better is the fit of the model to the data. In this case it would be best to try alternative models and opt for the one that produces largest *P* values.

  No model has an exact fit. With enough data the goodness of fit test will always reject the model. However, what one is looking for is whether the model is good enough for analysis purposes. One can still make inferences from a model not fitting well but caution is needed.

5). **Table of Observed and expected frequencies**. How well the model fits the data is next assessed by comparing observed and expected frequencies by estimated probabilities from lowest to highest. Group 1 contains data with lowest estimated probability; group 8 contains data with highest estimated probability. There is disparity at the levels of Groups 7 and 8 for both Thall = 1 and Thall = 0.

6). **Measures of Association**. The table displays the numbers and percent of concordant, discordant and tied pairs, as well as the common correlation tests. These values measure the association between observed responses and predicted probabilities. There were 34 individuals with +ve and 66 who had –ve results on thallium test. So in all there are 34 x 66 = 2244 pairs with different values. A pair is concordant if a subject with +ve thallium test has a higher probability of being positive, and discordant if negative. 74% of the pairs proved to be concordant. This can be used as a measure of prediction.

A series of tests under Summary measures consider concordance discordance as follows:

Somer's D is obtained by the following:

$$\frac{\textit{how many more concordant than discordant pairs exist}}{\textit{Total number of pairs}}$$

Goodman-Kruskal-Gamma is obtained by the following:

$$\frac{\textit{how many more concordant than discordant pairs}}{\textit{Total number of pairs} - \textit{number of ties}}$$

Kendall's Tau is obtained as follows:

$$\frac{\textit{how many more concordant than discordant pairs}}{\textit{Total number of pairs including pairs with same response}}$$

Large values for Somer's D, Goodman-Kruskal-Gamma and Kendall's Tau – a indicate that the model has a good predictive ability.

In our case there were 74% pairs concordant and 23% discordant giving 50% better chance for pairs to be concordant than discordant.

At the end a choice of diagnostic curves can be looked at. These are discussed under "Regression Diagnostics".

Comment

The main statistical work of logistic regression is finding the best fitting coefficients for α, β_1, β_2,…. β_p when a set of data is analysed. The fit of the model is evaluated, and the coefficients are appraised to assess the impact of the individual variables. The basic approach is the same as in multiple linear regression, but the mathematics is different.

Basic format for multiple logistic regression is that the outcome variable is binary, coded as 1= event; 0 = no event. The independent variables can be continuous, or categorical. The odds of the estimated probability are P ÷ (1 – P). The logistic model is formatted as the logit or natural logarithm of the odds. The logistic expression is

$$\text{Log}_e \{p/1\text{-}p\} = \alpha + \beta_1 X_1 + \beta_2 X_2 ... + \beta_p X_p \text{ for a full regression on all p variables.}$$

As mentioned above the logistic method can process any type of explanatory variables. The coefficients are interpreted as odds ratios. The explanatory variables act as 'predictors' or explainers of the target event.

Choosing the best model.

Decision about the best model is made using the log likelihood ratios. In this respect the log likelihood ratio in logistic regression is analogous to F and R^2 in linear regression. For the same set of data, two models with the same number of variables can be compared by using each model's value of log likelihood ratio. Inclusion of more variables would almost always improve the fit as it does also in linear regression. Therefore, when two models have different numbers of variables the comparison is not correct.

Suppose a researcher wishes to assess the importance of a particular risk factor. One way would be to perform two separate sets of analysis, one including the risk factor of interest (usually called the full model), and the other without it (hence called the reduced model). The computer output for each model would give the log-likelihood statistic. All that the researcher has to do is to simply work out the difference to obtain the likelihood statistic related to the risk factor of interest. The difference between log likelihood statistic for two models, one of which is a special case of the other, has an approximate chi-square distribution in large samples. The degree of freedom for this chi-square test in our example would be one, because one parameter has been added in the form of one additional risk factor. If we had added two risk factors to the model the degree of freedom would have been two, and so on. So the hypothesis that the additional risk factor makes no difference is tested by computing $\chi^2 = -2$ (log likelihood full model – log likelihood reduced model). Having obtained the value of χ^2 its significance level can be looked up by entering the table of chi-square at one degree of freedom.

A second approach that does not use the likelihood method is doing the **Wald test**. Recall that the Wald statistic is obtained by dividing the value of the β coefficient of the factor of interest by its standard error. For large samples the likelihood ratio and Wald^2 (Wald squared) give approximately the same value. If the study is large enough it does not matter which statistic is used. In the case of small and moderate sized samples the two statistics may give different results. In such a case the likelihood ratio statistic is better than Wald. So the rule of thumb is, when in doubt use likelihood ratio. Wald statistic is useful when one is fitting only one model viz. the full model.

Wald statistic also provides information whether the β coefficient of an explanatory variable (predictor) is significantly different from zero. It is analogous to the '*t*' test in multiple linear regression. As such it is making a significant contribution to the prediction of the outcome.

In interpreting the beta coefficients it has been pointed out that each coefficient is the natural logarithm of the odds ratio for the variable it represents. The odds ratio itself would be *e* raised to the power equal to the numerical value of the coefficient concerned. If the variable is categorical and coded as 1/ 0 the interpretation of the β coefficient becomes particularly interesting. Now we can calculate the probability of event when the particular risk factor is present or absent by first calculating the value of Y when β = 1 and when β = 0, and then substituting the two values of Y in the formula $P = e^y / (1+e^y)$.

Conditional and unconditional logistic regression

Until recently most widely used computer packages offered unconditional procedures. Now several include facilities for conditional estimates also. In deciding between the two the researcher needs to consider the number of parameters in the model relative to the number of

subjects. Unconditional logistic regression is preferred if the number of parameters is small relative to the number of subjects. For example, consider a case-control study of 100 matched pairs making the number of subjects 200. Because of matching 99 dummy variables would be created to represent each pair. Add to this the intercept and the risk factor of interest. The number of parameters would be a minimum of 103. This situation requires conditional logistic regression. A simple rule of thumb is use conditional logistic regression if matching has been done, and unconditional if there has been no matching. A second rule of thumb is when in doubt use conditional because it always gives unbiased results. The unconditional method is said to overestimate the odds ratio if it is not appropriate.

Regression Diagnostics

In the case of multiple linear regression diagnostics were based on residuals. Recall that there are two types of residuals – raw residuals, which are the differences between the observed and fitted values of the outcome variables, and standardised residuals.

In the case of logistic regression Residual = 1 – (estimated probability). The software calculates residuals for each case and can standardise them to help evaluate the fit of the model. Several schemes are available for standardisation, and different computer packages use different methods. The common types of residuals are:

Standardised Pearson Residuals. These are similar to the standardised residuals of the multiple linear regression.

Deviance residuals. Residual deviance in logistic regression corresponds to the residual sum of squares in multiple linear regression. Standardised deviance residuals are often used to detect extreme or outlying observations.

The **Delta-beta** figure shows how many observations are exercising undue influence on the β coefficients.

A fourth type of residual is called **deletion residual**, and is based on the change observed when a particular observation is excluded from the data set. Some authors call them likelihood residuals.

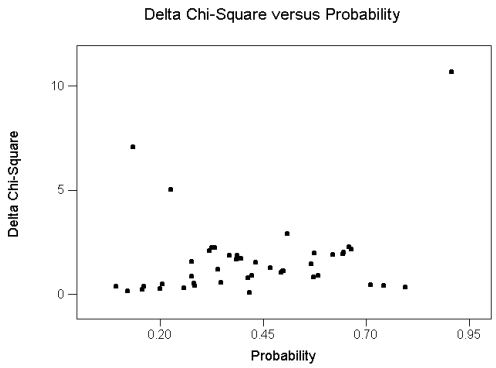
In all cases, large residuals indicate observations poorly fitted by the model.

We now consider the diagnostic graphs available in MINITAB. These are in two groups:

- Those involving event probability
- Those involving leverage

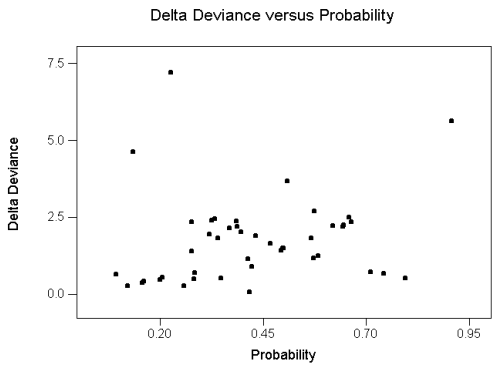
Diagnostic plots involving event probability

Delta chi square vs. probability



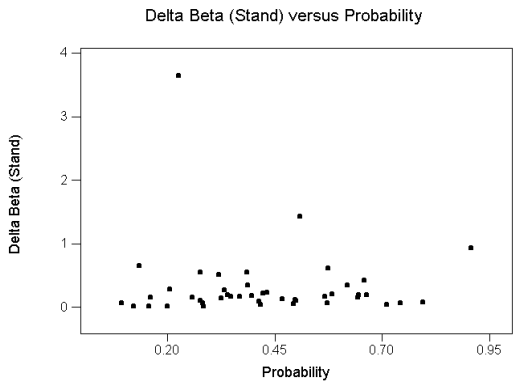
This plot helps to identify patterns that did not fit well. Delta Chi square measures change in the Pearson goodness-of-fit statistics due to deleting a particular factor. Points that are away from the main cluster correspond to low predicted probability. In the graph above 3 data points appear to be far away from the rest.

Delta Deviance vs. probability



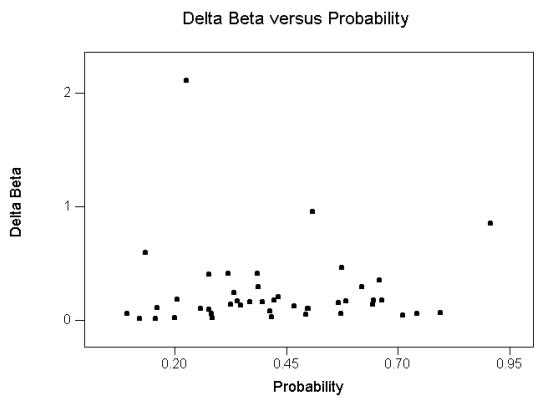
This plot helps to identify observations that have been fit well by the model. Delta deviance measures change in the deviance goodness-of-fit statistic due to deleting a particular observation. Large delta deviance values correspond to low predicted probabilities.

Delta beta (standardized) versus probability.



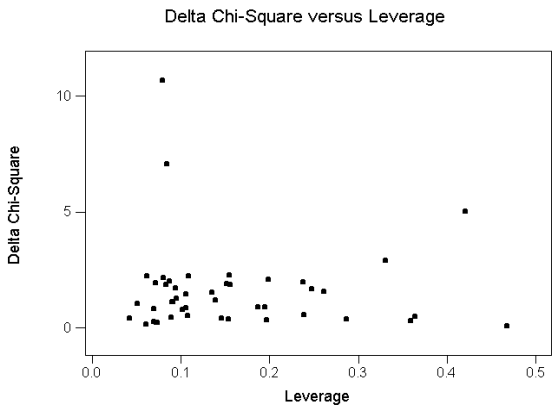
This plot identifies observations that have a strong influence on the estimated regression coefficients. A large delta beta (standardized) corresponds to a large leverage and/or a large residual.

Delta beta versus probability.



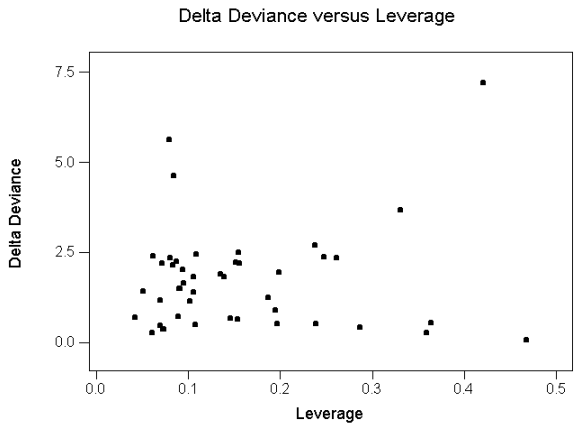
This plot identifies observations that have a strong influence on the estimated regression coefficients. Delta-beta is a measure of change in the regression coefficients due to deleting a particular observation.

Plots involving leverage



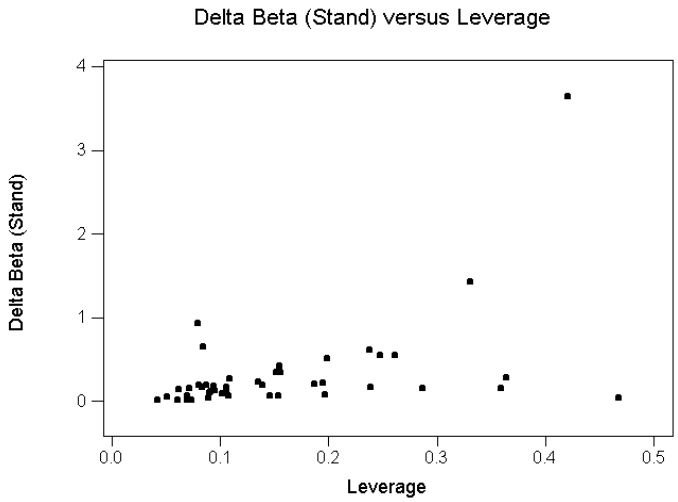
This plot identifies observations that have not been fit well in the model.

Delta deviance versus leverage



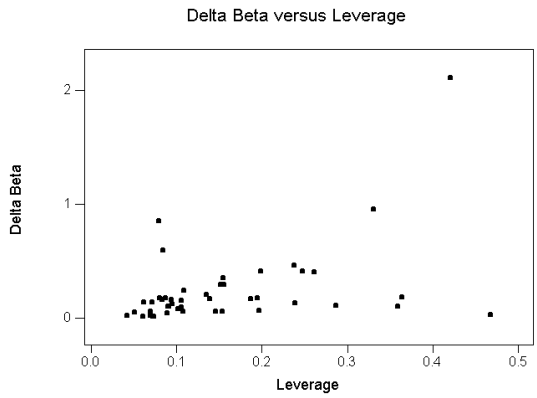
Delta deviance measures change in the deviance goodness-of-fit statistic due to deleting a particular observation.

Delta Beta (standardized) versus leverage



The plot identifies observations that have a strong influence on the estimated regression coefficients. Delta beta (standardized) measures change in regression coefficients due to deleting a particular observation.

Delta Beta versus leverage.



The plot identifies observations that have a strong influence on the regression coefficients. A large delta-beta often corresponds to an observation with a large leverage and /or large residual.

Practical Issues

- Extremely high values for parameter estimates (i.e. α and β coefficients) and for standard errors indicate that the number of explanatory variables is too large relative to the number of subjects. The solution is to either increase the number of subjects or remove one or more explanatory variables from the analysis.

- Like multiple linear regression, logistic regression also is sensitive to collinearity among the explanatory variables. The telltale sign of collinearity is high values for standard errors in parameter estimates.

- It is not uncommon for one or more cases to be quite poorly predicted. A case that in reality is in one category of outcome, may be shown to have a high probability of belonging in another category. If there are several such cases then obviously the model is not fitting the data well. Diagnostic plots are useful for identifying the outlying cases.