# Statistical Model Building & Logistic Regression

*Mr. Kannan Mahadevan, Biostatistician, Laico*

## Introduction

In a sequel to the previous article on straight line regression, we will focus on the special type of multiple linear regression- Logistic regression. If the goal of research is find out the causal relationship between the dependent and independent variables, the regression methods are more useful. The straight line regression can be extended to class of multiple regression analysis. The Multiple regression is useful method to characterize the relationship between the suppose to causal factors and effect describing the strength of the relationship through quantitative formula. Before pitching into the topic of interest, let us digress briefly about the statistical model building process pertinent to the regression modelling in general.

## Principles of statistical modelling

### a) Exploratory data analysis

Any analysis of data should begin with a consideration of each variable separately, both to check on data quality (for example, are the values plausible?) and to help with model formulation.

1. What is the scale of measurement? Is it continuous or categorical? If it is categorical how many categories does it have and are they nominal or ordinal?
2. What is the shape of the distribution? This can be examined using frequency tables, dot plots, histograms and other graphical methods.
3. How is it associated with other variables? Cross tabulations for categorical variables, scatter plots for continuous variables, side-by-side box plots for continuous scale measurements grouped according to the factor levels of a categorical variable, and other such summaries can help to identify patterns of association. For example, do the points on a scatter plot suggest linear or non-linear relationships? Do the group means increase or decrease consistently with an ordinal variable defining the groups?

### b) Model formulation

The models in this article (that eventually follows) involve a single response variable *Y* and usually several explanatory variables ($X_i$ 's). Knowledge of the context in which the data were obtained, including the substantive questions of interest, theoretical relationships among the variables, the study design and results of the exploratory data analysis can all be used to help formulate a model. Of course, the model formulation has two components:

1. Probability distribution of *Y* the response variable, for example, $Y \sim N(\mu, \sigma^2)$.
2. Equation linking the Mean value (expected value) of *Y* with a linear combination of the explanatory variables, for example, $E(Y) = \alpha + \beta x$
   (Note: E(Y) = mean value of Y)

### c) Parameter estimation

The most commonly used estimation methods in regression are maximum likelihood estimator. This method has many desirable properties. The parameter estimated has desirable properties statistical inference which will follow the model building. Besides this method, there are other methods currently packaged with all major software.

### d) Model Checking

After estimating the parameter one has to check the assumptions and assess the distribution of the predictor variables relevant to the methods used. Let me give brief overview of the statistical techniques used in the process, I would like to elucidate about one particular tool known as residuals in regression diagnostics.

Firstly, consider a model involving the normal distribution for the outcome variables. Residuals, are

important tools for checking the assumptions made in formulating a model. They should usually be independent and have a distribution which is approximately normal with a mean of zero and constant variance. They are also expected to be unrelated to the explanatory variables. These residuals, after some manipulations are standardized. These standardized residuals can be compared to the Normal distribution to assess the adequacy of the distributional assumptions and to identify any unusual values. This can be done by inspecting their frequency distribution and looking for values beyond the likely range; for example, no more than 5% should be less than "1.96 or greater than +1.96 and no more than 1% should be beyond ± 2.58.

The standardized residuals plotted against each of the explanatory variables that are included in the model presents predictive power of the model graphically. If the model adequately describes the effect of the variable, there should be no apparent pattern in the plot. If it is inadequate, the points may display curvature or some other systematic pattern which would suggest that additional or alternative terms may need to be included in the model. The residuals should also be plotted against other potential explanatory variables that are not in the model. If there is any systematic pattern, this suggests that additional variables should be included.

#### *e)* Inference and interpretation

It is sometimes useful to think of scientific data as measurements composed of a message, or **signal,** that is distorted by **noise.** For example: Suppose there is evidence based on the collected information that birth weight of the babies increase with gestational age, we would like to know if it is same among both genders. The 'signal' is the birth weight of the babies and the 'noise' comes from all the genetic and environmental factors that lead to individual variation. A goal of statistical modelling is to extract as much information as possible about the signal. In practice, this has to be balanced against other criteria such as simplicity. Accordingly a simpler or more parsimonious model that describes the data adequately is preferable to a more complicated one which leaves little of the variability 'unexplained'. To determine a parsimonious model consistent with the data, we test hypotheses about the parameters. Hypothesis testing is performed in the context of model fitting by defining a series of nested models corresponding to different hypotheses. Then the question about whether the data support a particular hypothesis can be formulated in terms of the adequacy of fit of the corresponding model relative to other more complicated models.

## Multiple Linear regression Methods - Logistic regression

### Introduction

Logistic regression is a form of statistical modelling that is often appropriate for categorical outcome variables. It describes the relationship between a categorical response variable and set of explanatory variables. The response variable usually has two outcomes, but it may be polychotomous (more than two outcomes) that is have more than two response levels. They can be either nominally or ordinally scaled. This article will mainly address the use and necessity of logistic regression when the response variable has dichotomous nature.

Historically, one of the first uses of regression-like models for binomial outcome data was for bioassay results (Finney, 1973). Responses were the proportions or percentages of 'successes'; for example, the proportion of experimental animals killed by various dose levels of a toxic substance. Such data are sometimes called **quantal responses**. The aim is to describe the probability of 'success', $\partial$, as a function of the dose, $x$; for example, $g(\pi) = \beta_1 + \beta_2$ x.

Consider for example, *Beetle mortality,* The Table1 shows numbers of beetles dead after five hours exposure to gaseous carbon disulphide at various concentrations (data from Bliss, 1935). Figure1 shows the proportions $p_i = y_i/n_i$ plotted against dose $x_i$ (actually $x_i$ is the logarithm of the quantity of carbon disulphide). We want to pick model that is well suited to modelling a probability, since the probability ranges from 0 to 1 as dosage varies can take any nonnegative value. A mathematical function named logistic function is well suited to modelling a

probability for this data. The characteristic of this function are a) whatever the value of the explanatory variables the model will be predicting the risk of disease. b) Another reason why the logistic model is so popular relates to the sigmoid shape of the logistic function and as sigmoid shape applies to the variety of disease conditions. The proportion of killed beetles is modelled through the logistic function. In this example, proportion of killed is dependent variable and dosage administered is independent variable.

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)},$$

for $x_i$ th dosage where i=1,..,n. After little mathematical work of the above equation, one can obtain logarithm of odds with the given dose level

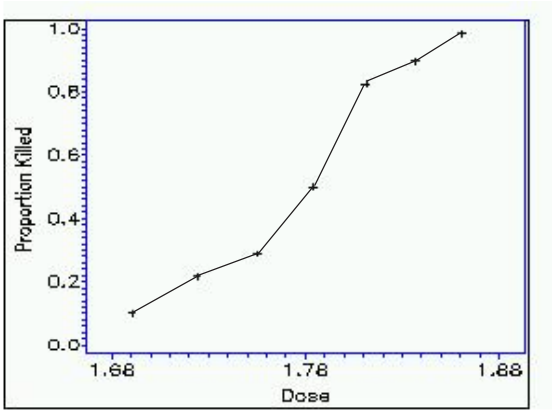
$$x_i \ \log \frac{\pi_i}{(1-\pi_i)} = (\beta_1 + \beta_2 x_i) \text{ for } i = 1,..,n.$$

Note that right hand side of the equation is linear with respect to the $x_i$. The logarithm of odds is commonly known as logit.

In epidemiologic studies, the logistic function is used to state individual's risk of developing a disease. One advantage is that whatever the value of the explanatory variables the model will be predictive the risk of disease. Another reason why the logistic model is so popular relates to the sigmoid shape of the logistic function (See figure1 for typical sigmoid shape) as sigmoid shape applies to the variety of disease conditions. In this example, the probability of getting killed increases as dosage level is increased. After rising over a range of intermediate values of dose, it remains close 1 once $x_i$ gets large enough.

*Table1: Beetle mortality data*

| Dose, xi ($Log_{10}CS_2$mgl"1) | Number of beetles, ni | Number of killed, yi |
|---|---|---|
| 1.6907 | 59 | 6 |
| 1.7242 | 60 | 13 |
| 1.7552 | 62 | 18 |
| 1.7842 | 56 | 28 |
| 1.8113 | 63 | 52 |
| 1.8369 | 59 | 53 |
| 1.861 | 62 | 61 |
| 1.8839 | 60 | 60 |




## Odds Ratio - measure of association:

The regression coefficients $\beta_j$ in the logistic model play an important role in providing information about the relationships of the predictors in the model to the dependent variable. For the logistic model, quantification of these relationships involves a parameter called odds ratio.

Let D be an out come of an event and define odds (D) = $\frac{pr(D)}{1- pr(D)}$

Consider an example, where lung cancer status is determined only by smoking status. For the time let us assume that data satisfactorily follows assumptions and logistic model is fit, say

log Pr(Y=1) = ($\beta_1$+$\beta_2$(smoking status), where Y denotes lung cancer status (1= yes, 0=no). Therefore the odd ratio for smokers versus non-smokers denoted by O.R as $_{S \text{ vs. } NS} = \frac{\text{odds } (S)}{\text{odds } (NS)}$

Note that , substitution and mathematical manipulation on mathemetical formula yields, Odds of smokers and Odds of non-smokers

( log Pr(Y=1) | Smokers) → log [odds (Smokers)] = $\beta_1$+$\beta_2$ → log odds (Smokers) = exp ( $\beta_1$+$\beta_2$ )

( log Pr(Y=1) | Non Smokers) → log [odds (Non Smokers)] = $\beta_1$+$\beta_2$ → log odds (Non Smokers) = exp ( $\beta_1$+$\beta_2$ )

rewriting odds ratio O.R $_{S \text{ vs. } NS} = \frac{\exp(\beta_1+\beta_2)}{\exp \beta_1} \rightarrow \exp \beta_2$

In other words, for the above example involving dichotomous predictors, the odds ratio comparing the two categories of the predictor is obtained by exponentiating the coefficient of the predictor in the logistic model.

## Understanding Basics in logistic modelling:

Essentially given a data**,** the logistic model may be fit based on the clinical and biological relevance of the independent variables on the dependent variable. After fitting the model to the data, one may need to assess how well it fits the data, or how close the model predicted values are to the corresponding observed values. The test statistics that assess fit in this manner are known as **goodness-of-fit statistics**.

## Model Fitting & Goodness of fit statistic:

Generally, the logistic models are fit through the softwares which takes care of the complex mathematical iterations to churn out the parameters of the model. In fact, all software have inbuilt functions to estimate the parameters through different methods.The goodness of fit of the data is useful to assess the model fit to the data.They address the differences between observed and predicted values or their ratio, in some appropriate manner. Of course, departure of the predicted values or proportions from the observed one should be essentially random. The statistics for testing goodness of fit have an approximate chi-square distributions when the sample is large for combination of the categories in the independent variables. The two traditional goodness-of-fit test statistics used are the Pearson chi-square, $Q_p$, and the likelihood ratio chi-square, $Q_{L,}$ also known as the deviance. If they are larger than tolerable values, one may have an over simplified model and you need to identify some other factors to better explain the variation in the data. Sometimes a logistic model is considered reasonable, but the goodness-of-fit test statistics indicate that too much variation remains (usually the deviance is examined). This condition is known as over dispersion. There are methods available to account for the over dispersion in the model.

## An example:

The following data are based on a study on coronary artery disease by Koch, Imrey, et al. 1985. Investigators were interested in whether electrocardiogram (ECG) measurement was associated with disease status. Gender was thought to be confounder for the disease, so data were post stratified data into male and female groups.

*Table1.1 Coronary artery disease data*

| Sex | ECG Segment Depression | Coronary artery disease | |
|---|---|---|---|
| | | Yes | No |
| Female | <0.1 | 11 | 4 |
| Female | >=0.1 | 10 | 8 |
| Male | <0.1 | 9 | 9 |
| Male | >=0.1 | 6 | 21 |

The table 1.1 describes the data collected on the patients who underwent diagnosis. It is easy to understand the diagnosis information from the table. The first row can be rephrased as people who were all female and segment depression less than 0.1, there were 11 people who had coronary artery disease and 4 did not have coronary artery disease.

An usual model for the data is the one that includes effects for Sex and ECG. That is to find out if the sex and ECG affects the occurance of coronary artery disease. When all the supposed predictor variables are included in the model without product of them and square of the variables, it is called main effects model. The main effects model tests the effects of each supposed predictors of the disease statistically. Mathematically logistic model including the main effects can be described as in the figure 1.

$$\begin{bmatrix} \log it(\theta_{11}) = \alpha \\ \log it(\theta_{12}) = \alpha + ..... + \beta_2 \\ \log it(\theta_{21}) = \alpha + \beta_1 + \\ \log it(\theta_{22}) = \alpha + \beta_1 + \beta_2 \end{bmatrix}$$

*Figure-Mathematical equation of the logistic model*

Here è $_{11}$ represents probability of coronary artery disease in the female population who have segment

depression less than 0.1.Taking logarithm of the above equations one may obtain the table 1.2 and their respective odds. The quantity á is the log odds of coronary artery disease for females with an ECG of less than 0.1 segment depression. The parameter $\hat{a}_1$ is the increment in log odds for males, and $\hat{a}_2$ is the increment in log odds for having an ECG of atleast 0.1 segment depression. The above mathematical equation can be interpreted as in the following table.

*Table1.2- Odds of CA Disease.*

| **Sex** | **ECG** | **Pr(CA disease) = $\theta_{hi}$** | **Odds of CA Disease** |
|---|---|---|---|
| Female | <0.1 | $e^{\alpha}$/ 1+$e^{\alpha}$ | $e^{\alpha}$ |
| Female | >=0.1 | $e^{\alpha+\beta 2}$/1+$e^{\alpha+\beta 2}$ | $e^{\alpha+\beta 2}$ |
| Male | <0.1 | $e^{\alpha+\beta 1}$ /1+$e^{\alpha+\beta 1}$ | $e^{\alpha+\beta 1}$ |
| Male | >=0.1 | $e^{\alpha+\beta 1+\beta 2}$/1+$e^{\alpha+\beta 1+\beta 2}$ | $e^{\alpha+\beta 1+\beta 2}$ |

When the analysis was done using the SAS software the following were obtained among the output. Based on this output, the parameter estimates and other useful information are extracted, simultaneous significance of the predictors are tested and goodness of fit of the model is validated. First let us explore the goodness of fit statistic provided in the Table 2.1. The Deviance and Pearson statistic are evaluated for the significance of the parameter in the model. A p-value greater than 0.05 suggests that model fits the adequately.

*Table 2.1 Goodness of fit of the model to the data.*

| Deviance and Pearson Goodness-of-Fit Statistics | | | | |
|---|---|---|---|---|
| Criterion | DF | Value | Value/DF | Pr > ChiSq |
| Deviance | 1 | 0.2141 | 0.2141 | 0.6436 |
| Pearson | 1 | 0.2155 | 0.2155 | 0.6425 |

As the model fits the data adequately, it is appropriate to examine the parameter estimates from the model (Table 2.2). The p-values of the parameters are less than the significant level of 0.05 suggesting the population values are not equal to zero. The variable sex and ECG are significant compared to a significance level of 0.05. So the results of the model predicted odds of coronary disease is listed in the Table 2.2- 2.3.

*Table 2.2 Analysis of Maximum Likelihood Estimates*

| Parameter | Estimate | Standard Error | Interpretation |
|---|---|---|---|
| A | -1.1747 | 0.4854 | log of odds of coronary disease for females with ecg <0.1 |
| sex | 1.2770 | 0.4980 | Increment in log of odds for males |
| ecg | 1.0545 | 0.4980 | Increment in log of odds for high ecg |

*Table 2.3 Model predicted logits and odds of CA disease.*

| Sex | ECG | Pr(CA disease) = è$_{hi}$ | Odds of CA Disease |
|---|---|---|---|
| Female | <0.1 | $\hat{\alpha}$ = -1.1747 | $e^{\hat{\alpha}}$ = 0.3089 |
| Female | >=0.1 | $\hat{\alpha}+\hat{\beta}_2$ = -0.1202 | $e^{\hat{\alpha}+\hat{\beta}_2}$ = 0.8867 |
| Male | <0.1 | $\hat{\alpha}+\hat{\beta}_2$ = 0.1023 | $e^{\hat{\alpha}+\hat{\beta}_1}$ = 1.1077 |
| Male | >=0.1 | $\hat{\alpha}+\hat{\beta}_1+\hat{\beta}_2$ = 1.1568 | $e^{\hat{\alpha}+\hat{\beta}_1+\hat{\beta}_2}$ = 3.1797 |

## Conclusion:

In the present paper, we did talk about the basic aspects of the statistical modelling and the usefulness of the data. The statistical data modelling especially regression in reality is very much challenging presenting abnormal data that may not be discerned very well. Somtimes the model may not fit the data adequately. In those cases we need to find alternative method that will do the intended analysis in line with

the goal of the research. The successful strategy lies at the design of studies and selecting appropriate statistical methods that will have robustness to the data anamolies. It is better to anticipate the problems of data anomalies before we find it difficult to handle it. However with the plethora of methods and techniques to measure the association patterns the onus rests on us to choose more than one technique if there is snag with a method.

## References


1. *Alan Agresti 1996. Categorical data analysis. 2nd Edition- Wiley.*
2. *Kleinbaum, Kupper, Muller and Nizam 1998.Applied regression analysis and other multivariable methods. 3rd Edition –. Duxbury press, CA.*
3. *Draper, N.R. and Smith, H. (1998), Applied Regression Analysis (3rd ed), Wiley. Applied Regression Analysis*