Data Champion module

Lean Six Sigma Master Black Belt

INDIA
USA
CHINA
SINGAPORE

QAI

QAI GLOBAL INSTITUTE

Logistic Regression

- Logistic or "Logit" regression investigates the relationship between response variables (Y's) and one or more predictor variables (X's) where:
  - **Y's are categorical, not continuous**

  - X's can be either continuous or categorical

QAI

Logistic Regression

- Both logistic regression and least squares regression investigate the relationship between a response variable and one or more predictors.
- A practical difference between them is that logistic regression techniques are used with categorical response variables, and linear regression techniques are used with continuous response variables.

MINITAB provides three logistic regression procedures that you can use to assess the relationship between one or more predictor variables and a categorical response variable of the following types:

Variable type	Number of categories	Characteristics	Examples
Binary	2	two levels	success, failure yes, no
Ordinal	3 or more	natural ordering of the levels	none, mild, severe fine, medium, coarse
Nominal	3 or more	no natural ordering of the levels	blue, black, red, yellow sunny, rainy, cloudy

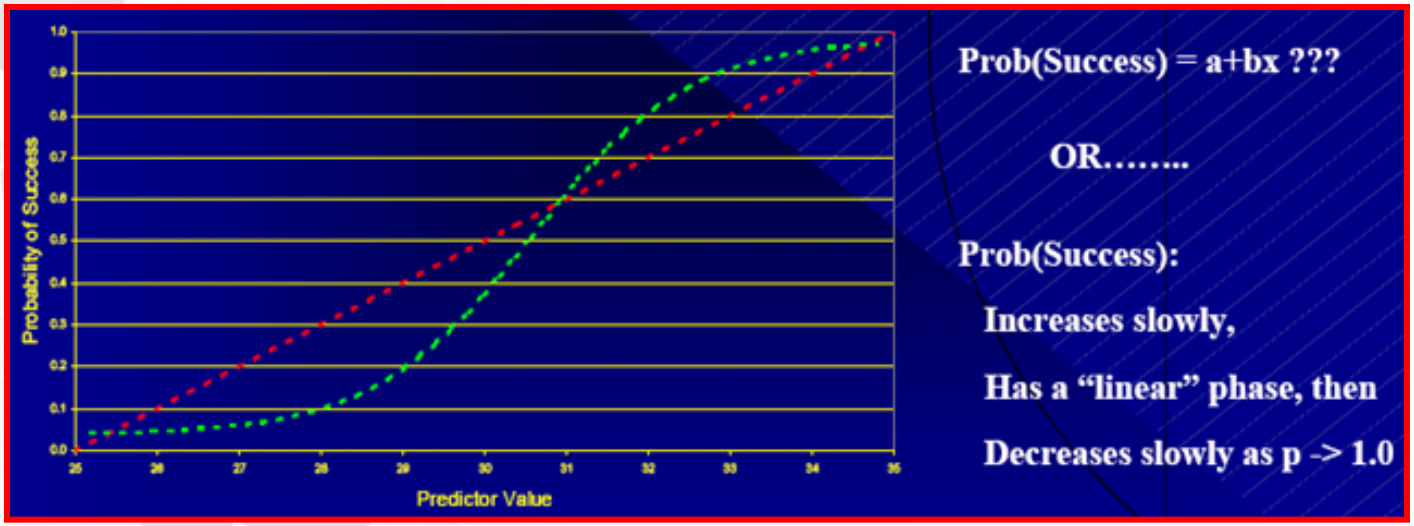
Logistic Regression and Minitab

- The logistic procedures in Minitab can fit up to
  - 9 categorical inputs (factors), and
  - 50 continuous inputs (covariates)

QAI

Why Logistic Regression?

- Binomial data violates normality, equal variance assumption
  - $\mu_k = n * p$ $\sigma_k^2 = n * p * (1-p) = \mu_k * (1-p)$
  - Variance changes as the mean changes
  - The relationship between p, the likelihood of "success" and the predictor variables might not be linear



Introducing Odds

- Odds, O = p/(1-p) = Probability (event occurs)/Probability (Event does not occur)
  p = O/(O+1)

- Odds Example: Define an event as “an account paid in 30 days”
  - If the odds are defined as 6 to 1, this implies:
    - Pr (account paid) = 6/(6+1) = 6/7
    - Pr (account not paid) = (1/6)/(1/6+1) = 1/(1+6) = 1/7
    - Odds of account being paid = 6/1 = 6
    - Odds of account not being paid = 1/6

QAI

Introducing Odds Ratio

- Odds Ratio
  - Let the odds of the 1st shift completing their schedule be 5 to 2,
  - Let the odds of the 2nd shift completing their schedule be 10 to 1,
  - Shift (2)'s odds of completion (10) is 4 times larger than Shift (1)'s odds (2.5)
    - Ratio of Odds or "Odds ratio" of Shift (2) to Shift (1) = 10/2.5 = 4

Binary Logistic Regression

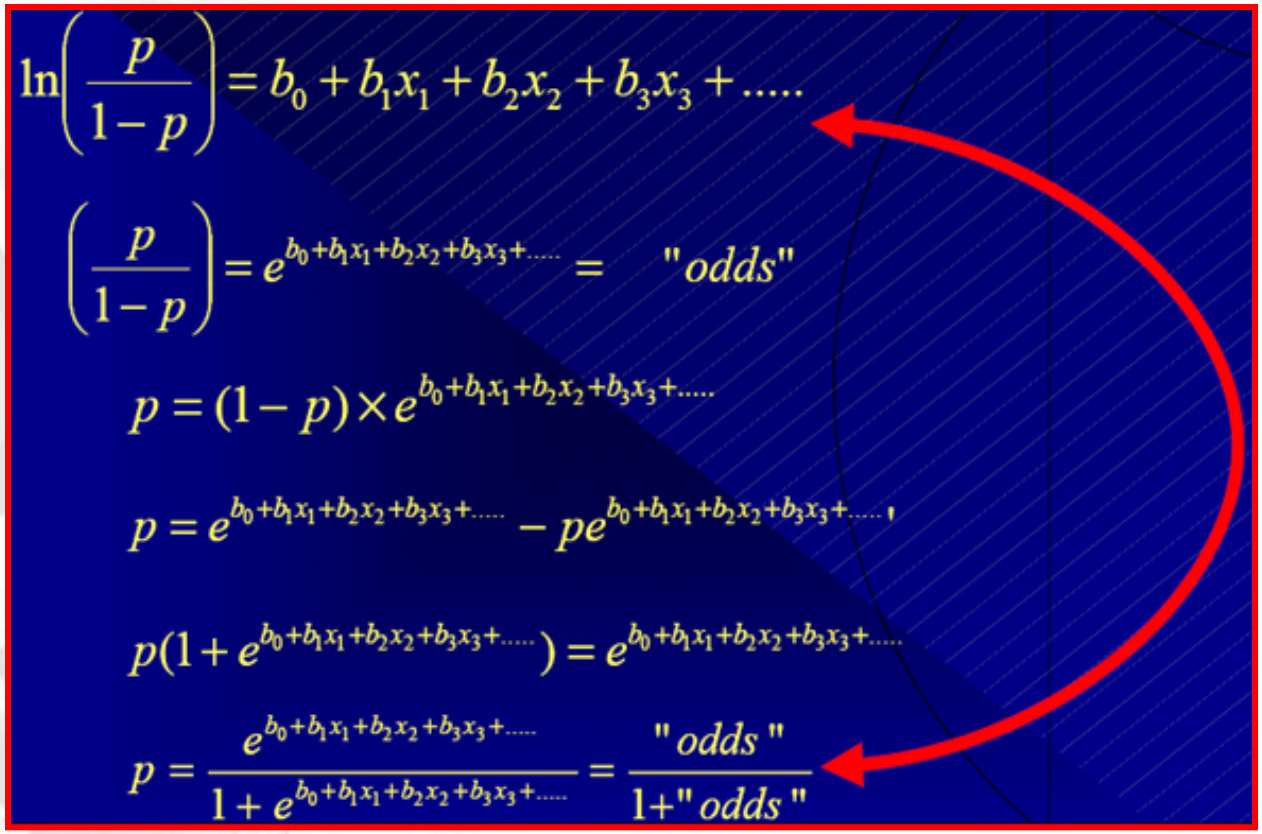
- We will demystify logistic regression using the simplest logistic regression – binary logistic regression (where the Y has just two potential outcomes, i.e., "yes" or "no," or 0 or 1)
- These events are often described as success or failure
- For each possible values for the independent (X) variables, there is a probability that a “success” occurs

The linear logistic model fitted by maximum likelihood is:

- $Y = b_0 + b_1*X_1 + b_2*X_2 \ldots\ldots + b_k*X_k$
- Where Y = logit transformation of the odds based on p = Prob(event)
  - Odds = $\left(\frac{p}{1-p}\right)$ Logit = $\ln\left(\frac{p}{1-p}\right)$

QAI

Deriving Probability from Logit Results



QAI

Requirements for Applying Binary Logistic Regression

- A dependent (Y) variable with two outcomes, such as pass/fail
  - One of the outcomes is defined as an **Event**

- One of more predictor (X) variables
  - Categorical X's are known as **factors**
  - Continuous X's are known as **covariates**
  - Factors will have two or more levels
  - One level is known as the **reference level**

Minitab will treat each factor as covariate unless it is specified as a factor

- Data which relates the level of the X variable(s) to the corresponding Y response
  - **All possible Y outcomes should be represented for EACH factor level (i.e. both pass and fail responses should be observed for ALL factor level(s)**

QAI

How Minitab defines Events and Reference Levels

- **Event (Y)**
  - If the response is numeric or date/time, Minitab chooses the highest value
    - For Y = 1 or 2, 2 will be defined as the "event"
  - If the response is text, Minitab chooses the name which is last in the ascending sort order
    - For two responses **pass** and **fail**, **pass** will be defined as the event

- **Reference Level for Factors (X)**
  - If the levels are all numeric or date/time, the reference level is the lowest numerical value
  - If the levels are text, the first level in the sort order is the reference level
    - For a two factor resulting in **new** or **old**, **new** will be the reference level because it is the lowest in the sort order

QAI

Statistical Perspective

- When the X variable is at the reference level, the probability that the event will occur or not occur can be examined in terms of “Odds”

- When the X variable changes from the reference level to another level, the odds ratio, or the fraction of change to the odds will be examined

- Logistical Regression uses a “Link Function”
  - A link function is a function used to fit the logistic regression model
  - The link function that will be used is the logit:
    - ln(p/1-p), or the natural logarithm of the Odds
  - Minitab offers other link functions

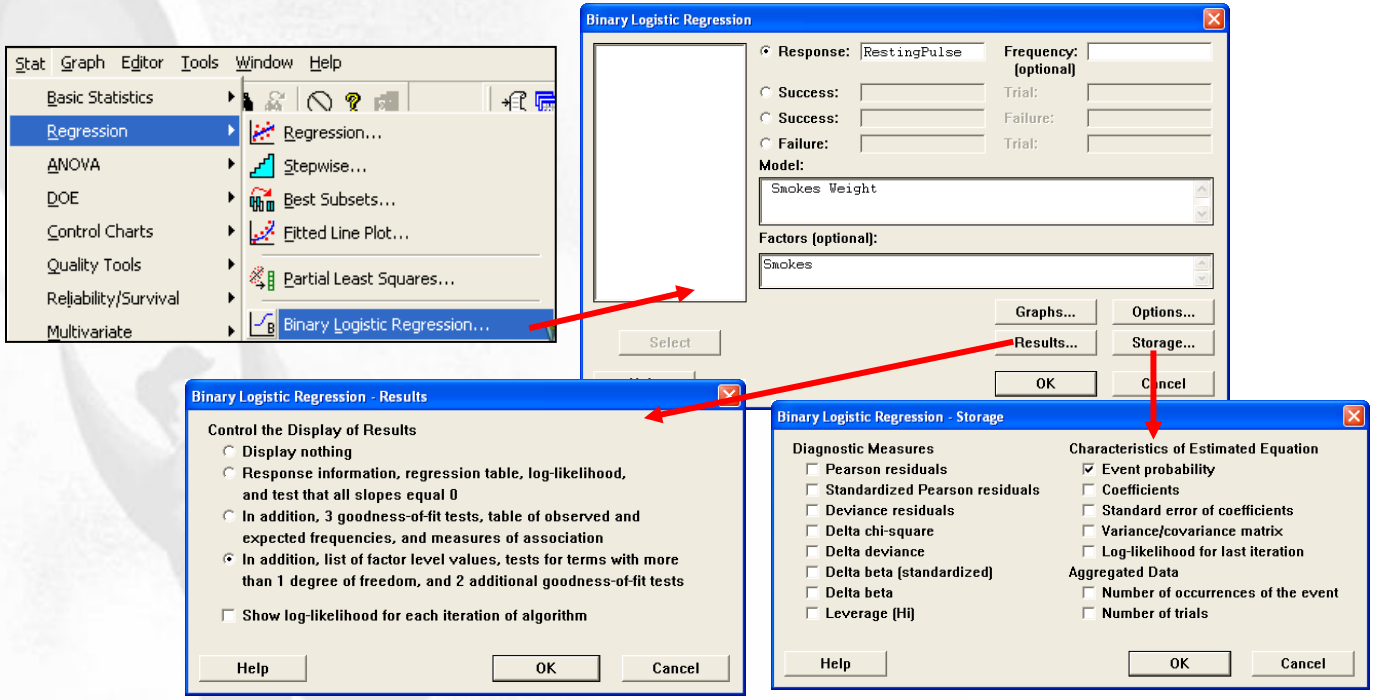
QAI

Minitab Example

- Consider the smokers' data set in LogisticRegression worksheet

P = Probability(Low Resting Pulse), examine:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \times [Smokes\ ?] + b_2 \times [Weight\ ]$$



Minitab Example – Output (Partial)

Binary Logistic Regression: RestingPulse versus Smokes, Weight

```
Link Function: Logit


Response Information

Variable      Value  Count
RestingPulse  Low       70  (Event)
              High      22
              Total     92


Factor Information

Factor  Levels  Values
Smokes       2  No, Yes


Logistic Regression Table

                                                        Odds      95% CI
Predictor       Coef    SE Coef       Z      P       Ratio   Lower   Upper
Constant    -1.98717    1.67930   -1.18  0.237
Smokes
 Yes        -1.19297   0.552980   -2.16  0.031        0.30    0.10    0.90
Weight     0.0250226  0.0122551    2.04  0.041        1.03    1.00    1.05

Log-Likelihood = -46.820
Test that all slopes are zero: G = 7.574, DF = 2, P-Value = 0.023
```

The function used to fit the model relating RestingPulse to Smokes and Weight is the Odds function (Logit)

The Event is RestingPulse Low

The factor Smokes has two levels No and Yes. The level No is the Reference Level

The p values for both Smokes and Weight are significant at the 0.05 alpha level

The Odds of a smoker having a low resting pulse is 0.3 that of a non-smoker having a low resting pulse. Thus, subjects who smoke tend to have a higher resting pulse

The Odds of a subject having a low resting pulse increases 1.03 times with each pound increase in weight. Although statistically significant, there is not much difference practically.

The Log –Likelihood term is displayed along with the statistic G. The G statistic refers to the null hypothesis that all coefficients associated with predictors are equal to 0 or not. The P Value < 0.05 reinforces the belief that there is a at least one coefficient that is significantly different from zero. Or we can also say that the p value of 0.023 means that either Smoke (factor) or Weight (covariate) or both will have a significant Odds ratio different from 1.

Minitab Example - Output

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	40.8477	47	0.724
Deviance	51.2008	47	0.312
Hosmer-Lemeshow	4.7451	8	0.784
Brown:			
General Alternative	0.9051	2	0.636
Symmetric Alternative	0.4627	1	0.496

Pearson and Deviance are both types of residuals for logistic models. They are useful measures for evaluating how well the selected model fits the data. The higher the p-value, the better the model fits the data.

For the data, both the Pearson and Deviance tests have p-values that are greater than 0.10 indicating that there is insufficient evidence for the model not fitting the data adequately when the α-level is less than or equal to 0.10.

The Hosmer-Lemeshow test assesses the model fit by comparing the observed and expected frequencies. The test groups the data by their estimated probabilities from lowest to highest, then performs a Chi-square test to determine if the observed and expected frequencies are significantly different.

Minitab Example - Output

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	9	10	Total
Low											
Obs	4	6	6	8	8	6	8	12	10	2	70
Exp	4.4	6.4	6.3	6.6	6.9	7.2	8.3	12.9	9.1	1.9	
High											
Obs	5	4	3	1	1	3	2	3	0	0	22
Exp	4.6	3.6	2.7	2.4	2.1	1.8	1.7	2.1	0.9	0.1	
Total	9	10	9	9	9	9	10	15	10	2	92

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

Pairs	Number	Percent	Summary Measures	
Concordant	1045	67.9	Somers' D	0.38
Discordant	461	29.9	Goodman-Kruskal Gamma	0.39
Ties	34	2.2	Kendall's Tau-a	0.14
Total	1540	100.0		

This table lets us see how well the data fits by comparing observed and expected frequencies

In this data there are 70 individuals with low resting pulse and 22 individuals with high resting pulse, resulting in 70*22=1540 pairs with different response values

Based on the model, the pair is concordant if the individual with a low pulse rate has a higher probability of having a low pulse rate, discordant if the opposite is true, and tied if the probabilities are equal

You can use these values as a comparative measure of prediction

Deriving Useful Interpretations

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \times [Smokes\ ?] + b_2 \times [Weight\ ]$$

$$\left(\frac{p}{1-p}\right) = e^{b_0 + b_1 \times [Smokes\ ?] + b_2 \times [Weight\ ]}$$

$$\left(\frac{p}{1-p}\right) = e^{b_0} e^{b_1 \times [Smokes\ ?]} e^{b_2 \times [Weight\ ]}$$

For each component, the Odds change by e^{bx}

QAI

Deriving Useful Interpretations (cont)

$$\left(\frac{p}{1-p}\right) = e^{b_0} e^{b_1 \times [Smokes\ ?]} e^{b_2 \times [Weight\ ]}$$

- When each X is at it's reference value:

$$\left(\frac{p}{1-p}\right) = e^{-1.987} = .1371$$

$$p = .1371 / (1 + .1371) = .1206$$

- Prob(Low Resting Pulse) at all reference levels = .1206

QAI

Deriving Useful Interpretations (cont)

$$\left(\frac{p}{1-p}\right) = e^{b_0} e^{b_1 \times [Smokes\ ?]} e^{b_2 \times [Weight\ ]}$$

- When a person smokes and all other X's are at their reference values:

$$\left(\frac{p}{1-p}\right) = e^{-1.987} e^{-1.1930} = e^{(-1.987-1.1930)} = e^{-3.18} = .0416$$

$$p = .0416 / (1 + .0416) = .04$$

- Prob(Low Resting Pulse and Smokes) = .04
- Change in odds = .0416/.1371 = .3034

QAI

Deriving Useful Interpretations (cont)

$$\left(\frac{p}{1-p}\right) = e^{b_0} e^{b_1 \times [Smokes\ ?]} e^{b_2 \times [Weight\ ]}$$

- When a person smokes and weighs 135 lbs:

$$\left(\frac{p}{1-p}\right) = e^{-1.987} e^{-1.1930} e^{(.02502 \times (135-95))}$$

Note: The 95 pounds in weight is the reference value for weight

$$= e^{(-1.987-1.1930+1.0008)} = e^{-2.1792} = .1131$$

$$p = .1131/(1+.1131) = .1016$$

- Prob(Low Resting Pulse when the patient Smokes at 135 lbs) = .1016
- Change in odds = .1131/.1371 = .825

QAI

Thank You