LOGISTIC REGRESSION

Nitin R Patel

Logistic regression extends the ideas of multiple linear regression to the situation where the dependent variable, y, is binary (for convenience we often code these values as 0 and 1). As with multiple linear regression the independent variables $x_1, x_2 \cdots x_k$ may be categorical or continuous variables or a mixture of these two types.

Let us take some examples to illustrate [1]:

Example 1: Market Research

The data in Table 1 were obtained in a survey conducted by AT & T in the US from a national sample of co-operating households. Interest was centered on the adoption of a new telecommunications service as it related to education, residential stability and income.

Table 1: Adoption of New Telephone Service

	High School or below		Some College or above	
	No Change in Residence during Last five years	Change in Residence during Last five years	No change in Residence during Last five years	Change in Residence during Last five years
Low Income	153/2160 = 0.071	226/1137 = 0.199	61/886 = 0.069	233/1091 = 0.214
High Income	147/1363 = 0.108	139/ 547 = 0.254	287/1925 = 0.149	382/1415 = 0.270

(For fractions in cells above, the numerator is the number of adopters out of the number in the denominator).

Note that the overall probability of adoption in the sample is 1628/10524 = 0.155. However, the adoption probability varies depending on the categorical independent variables education, residential stability and income. The lowest value is 0.069 for low- income no-residence-change households with some college education while the highest is 0.270 for high-income residence changers with some college education.

The standard multiple linear regression model is inappropriate to model this data for the following reasons:

1. The model's predicted probabilities could fall outside the range 0 to 1.

2. The dependent variable is not normally distributed. In fact a binomial model would be more appropriate. For example, if a cell total is 11 then this variable can take on only 12 distinct values $0, 1, 2 \cdots 11$. Think of the response of the households in a cell being determined by independent flips of a coin with, say, heads representing adoption with the probability of heads varying between cells.

3. If we consider the normal distribution as an approximation for the binomial model, the variance of the dependent variable is not constant across all cells: it will be higher for cells where the probability of adoption, p, is near 0.5 than where it is near 0 or 1. It will also increase with the total number of households, n, falling in the cell. The variance equals $n(p(1-p))$.

The logistic regression model was developed to account for all these difficulties. It has become very popular in describing choice behavior in econometrics and in modeling risk factors in epidemiology. In the context of choice behavior it can be shown to follow from the random utility theory developed by Manski [2] as an extension of the standard economic theory of consumer behavior.

In essence the consumer theory states that when faced with a set of choices a consumer makes a choice which has the highest utility (a numeric measure of worth with arbitrary zero and scale). It assumes that the consumer has a preference order on the list of choices that satisfies reasonable criteria such as transitivity. The preference order can depend on the individual (e.g. socioeconomic characteristics as in the Example 1 above) as well as attributes of the choice. The random utility model considers the utility of a choice to incorporate a random element. When we model the random element as coming from a "reasonable" distribution, we can logically derive the logistic model for predicting choice behavior.

If we let $y = 1$ represent choosing an option versus $y = 0$ for not choosing it, the logistic regression model stipulates:

$$\text{Probability}(Y = 1|x_1, x_2 \cdots x_k) = \frac{\exp(\beta_O + \beta_1 * x_1 + \cdots \beta_k * x_k)}{1 + \exp(\beta_O + \beta_1 * x_1 + \cdots \beta_k * x_k)}$$

where $\beta_0, \beta_1, \beta_2 \cdots \beta_k$ are unknown constants analogous to the multiple linear regression model.

The independent variables for our model would be:

$x_1 \equiv$ (Education: High School or below = 0, Some College or above = 1

$x_2 \equiv$ (Residential Stability: No change over past five years = 0, Change over past five years = 1

$x_3 \equiv$ Income: Low = 0 High = 1

The data in Table 1 is shown below in the format typically required by regression programs.

x_1	x_2	x_3	# in sample	#adopters	# Non-adopters	Fraction adopters
0	0	0	2160	153	2007	.071
0	0	1	1363	147	1216	.108
0	1	0	1137	226	911	.199
0	1	1	547	139	408	.254
1	0	0	886	61	825	.069
1	1	0	1091	233	858	.214
1	0	1	1925	287	1638	.149
1	1	1	1415	382	1033	.270
			10524	1628	8896	1.000

The logistic model for this example is:

$$Prob(Y = 1|x_1, x_2, x_3) = \frac{\exp(\beta_0 + \beta_1 * x_l + \beta_2 * x_2 + \beta_3 * x_3)}{1 + \exp(\beta_0 + \beta_1 * x_l + \beta_2 * x_2 + \beta_3 * x_3)}.$$

We obtain a useful interpretation for the coefficients by noting that:

$$\begin{aligned}
\exp(\beta_0) &= \frac{Prob(Y = 1|x_1 = x_2 = x_3 = 0)}{Prob(Y = 0|x_1 = x_2 = x_3 = 0)} \\
&= \text{Odds of adopting in the base case } \ (x_1 = x_2 = x_3 = 0) \\
\exp(\beta_1) &= \frac{\text{Odds of adopting when } \ x_1 = 1, x_2 = x_3 = 0}{\text{Odds of adopting in the base case}} \\
\exp(\beta_2) &= \frac{\text{Odds of adopting when } \ x_2 = 1, x_1 = x_3 = 0}{\text{Odds of adopting in the base case}} \\
\exp(\beta_3) &= \frac{\text{Odds of adopting when } \ x_3 = 1, x_1 = x_2 = 0}{\text{Odds of adopting in the base case}}
\end{aligned}$$

The logistic model is multiplicative in odds in the following sense:

Odds of adopting for a given x_1, x_2, x_3

$$\begin{aligned}
&= \exp(\beta_0) * \exp(\beta_1 x_1) * \exp(\beta_2 x_2) * \exp(\beta_3 x_3) \\
&= \left\{ \begin{matrix} Odds \\ for \\ basecase \end{matrix} \right\} * \left\{ \begin{matrix} Factor \\ due \\ to\ x_1 \end{matrix} \right\} * \left\{ \begin{matrix} Factor \\ due \\ to\ x_2 \end{matrix} \right\} * \left\{ \begin{matrix} Factor \\ due \\ to\ x_3 \end{matrix} \right\}
\end{aligned}$$

If $x_1 = 1$ the odds of adoption get multiplied by the same factor for any given level of x_2 and x_3. Similarly the multiplicative factors for x_2 and x_3 do not vary with the levels of the remaining factors. The factor for a variable gives us the impact of the presence of that factor on the odds of adopting.

If $\beta_i = 0$, the presence of the corresponding factor has no effect (multiplication by one). If $\beta_i < 0$, presence of the factor reduces the odds (and the probability) of adoption, whereas if $\beta_i > 0$, presence of the factor increases the probability of adoption.

The computations required to produce these maximum likelihood estimates require iterations using a computer program. The output of a typical program is shown below:

					95% Conf. Intvl. for odds	
Variable	Coeff.	Std. Error	p-Value	Odds	Lower Limit	Upper Limit
Constant	-2.500	0.058	0.000	0.082	0.071	0.095
x_1	0.161	0.058	0.006	1.175	1.048	1.316
x_2	0.992	0.056	0.000	2.698	2.416	3.013
x_3	0.444	0.058	0.000	1.560	1.393	1.746

From the estimated values of the coefficients, we see that the estimated probability of adoption for a household with values x_1, x_2 and x_3 for the independent variables is:

$$Prob(Y = 1|x_1, x_2, x_3) = \frac{\exp(-2.500 + 0.161 * x_1 + 0.992 * x_2 + 0.444 * x_3)}{1 + \exp(-2.500 + 0.161 * x_1 + 0.992 * x_2 + 0.444 * x_3)}.$$

The estimated number of adopters from this model will be the total number of households with values x_1, x_2 and x_3 for the independent variables multiplied by the above probability.

The table below shows the estimated number of adopters for the various combinations of the independent variables.

x_1	x_2	x_3	# in sample	# adopters	Estimated (# adopters)	Fraction Adopters	Estimated $Prob(Y = l|x_1, x_2, x_3)$
0	0	0	2160	153	164	0.071	0.076
0	0	1	1363	147	155	0.108	0.113
0	1	0	1137	226	206	0.199	0.181
0	1	1	547	139	140	0.254	0.257
1	0	0	886	61	78	0.069	0.088
1	1	0	1091	233	225	0.214	0.206
1	0	1	1925	287	252	0.149	0.131
1	1	1	1415	382	408	0.270	0.289

In data mining applications we will have validation data that is a hold-out sample not used in fitting the model.

Let us suppose we have the following validation data consisting of 598 households:

x_1	x_2	x_3	# in validation sample	# adopters in validation sample	Estimated (# adopters)	Error (Estimate -Actal)	Absolute Value of Error
0	0	0	29	3	2.200	-0.800	0.800
0	0	1	23	7	2.610	-4.390	4.390
0	1	0	112	25	20.302	-4.698	4.698
0	1	1	143	27	36.705	9.705	9.705
1	0	0	27	2	2.374	0.374	0.374
1	1	0	54	12	11.145	-0.855	0.855
1	0	1	125	13	16.338	3.338	3.338
1	1	1	85	30	24.528	-5.472	5.472
Totals			598	119	116.202		

The total error is -2.8 adopters or a percentage error in estimating adopters of -2.8/119 = 2.3%.

The average percentage absolute error is

$$\frac{0.800 + 4.390 + 4.698 + 9.705 + 0.374 + 0.855 + 3.338 + 5.472}{119}$$

$$= .249 = 24.9\% \text{ adopters.}$$

The confusion matrix for households in the validation data for set is given below:

	Observed		
	Adopters	Non-adopters	Total
Predicted:			
Adopters	103	13	116
Non-adopters	16	466	482
Total	119	479	598

As with multiple linear regression we can build more complex models that reflect interactions between independent variables by including factors that are calculated from the interacting factors. For example if we felt that there is an interactive effect b etween x_1 and x_2 we would add an interaction term $x_4 = x_1 \times x_2$.

Example 2: Financial Conditions of Banks [2]

Table 2 gives data on a sample of banks. The second column records the judgment of an expert on the financial condition of each bank. The last two columns give the values of two commonly ratios commonly used in financial analysis of banks.

Table 2: Financial Conditions of Banks

Obs	Financial Condition* (y)	Total Loans & Leases/ Total Assets (x_1)	Total Expenses / Total Assets (x_2)
1	1	0.64	0.13
2	1	1.04	0.10
3	1	0.66	0.11
4	1	0.80	0.09
5	1	0.69	0.11
6	1	0.74	0.14
7	1	0.63	0.12
8	1	0.75	0.12
9	1	0.56	0.16
10	1	0.65	0.12
11	0	0.55	0.10
12	0	0.46	0.08
13	0	0.72	0.08
14	0	0.43	0.08
15	0	0.52	0.07
16	0	0.54	0.08
17	0	0.30	0.09
18	0	0.67	0.07
19	0	0.51	0.09
20	0	0.79	0.13

$$\begin{aligned} * \text{ Financial Condition} &= 1 \quad \text{for financially weak banks;} \\ &= 0 \quad \text{for financially strong banks.} \end{aligned}$$

Let us first consider a simple logistic regression model with just one independent variable. This is analogous to the simple linear regression model in which we fit a straight line to relate the dependent variable, y, to a single independent variable, x.

Let us construct a simple logistic regression model for classification of banks using the Total Loans & Leases to Total Assets ratio as the independent variable in our model. This model would have the following variables:

Dependent variable:

$$\begin{aligned} Y &= 1, \quad \text{if financially distressed,} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

Independent (or Explanatory) variable:

$$x_1 = \text{Total Loans \& Leases/Total Assets Ratio}$$

The equation relating the dependent variable with the explanatory variable is:

$$Prob(Y = 1|x_1) = \frac{\exp(\beta_0 + \beta_1 * x_l)}{1 + \exp(\beta_0 + \beta_1 * x_l)}$$

or, equivalently,

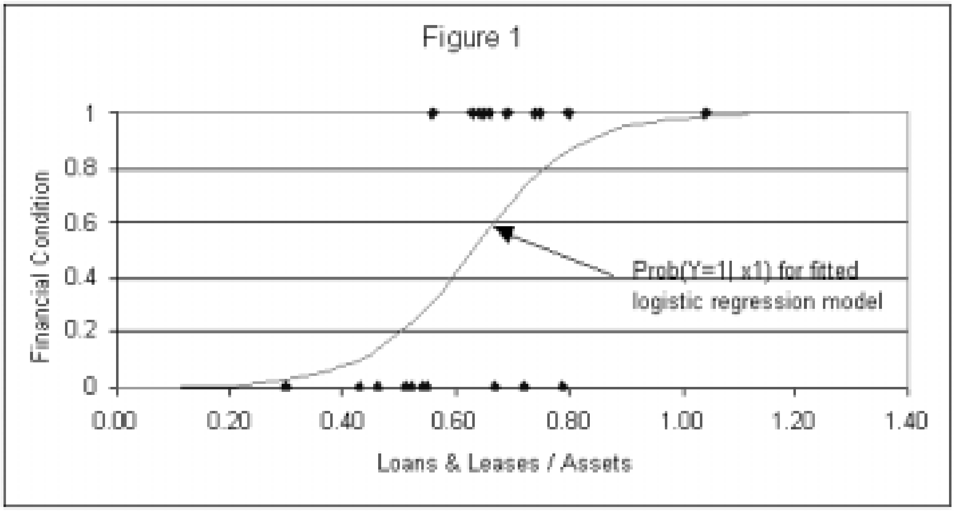
$$\text{Odds} \;\; (Y = 1 \text{ versus } \; Y = 0) = (\beta_0 + \beta_1 * x_l).$$

The Maximum Likelihood Estimates of the coefficients for the model are: $\hat{\beta}_0 = -6.926$, $\hat{\beta}_1 = 10.989$

So that the fitted model is:

$$Prob(Y = 1|x_1) = \frac{\exp(-6.926 + 10.989 * x_1)}{(1 + \exp(-6.926 + 10.989 * x_1)}.$$

Figure 1 displays the data points and the fitted logistic regression model.



We can think of the model as a multiplicative model of odds ratios as we did for Example 1. The odds that a bank with a Loan & Leases/Assets Ratio that is zero will be in financial distress = $\exp(-6.926) = 0.001$. These are the base case odds. The odds of distress for a bank with a ratio of 0.6 will increase by a multiplicative factor of $\exp(10.989*0.6) = 730$ over the base case, so the odds that such a bank will be in financial distress = 0.730.

Notice that there is a small difference in interpretation of the multiplicative factors for this example compared to Example 1. While the interpretation of the sign of β_i remains as before, its magnitude gives the amount by which the odds of $Y = 1$ against $Y = 0$ are changed for *a unit change* in x_i. If we construct a simple logistic regression model for classification of banks using the Total Expenses/ Total Assets ratio as the independent variable we would have the following variables:

Dependent variable:

$$\begin{aligned} Y &= 1, \text{ if financially distressed,} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

Independent (or Explanatory) variables:

$$x_2 = \text{Total Expenses/ Total Assets Ratio}$$

The equation relating the dependent variable with the explanatory variable is:

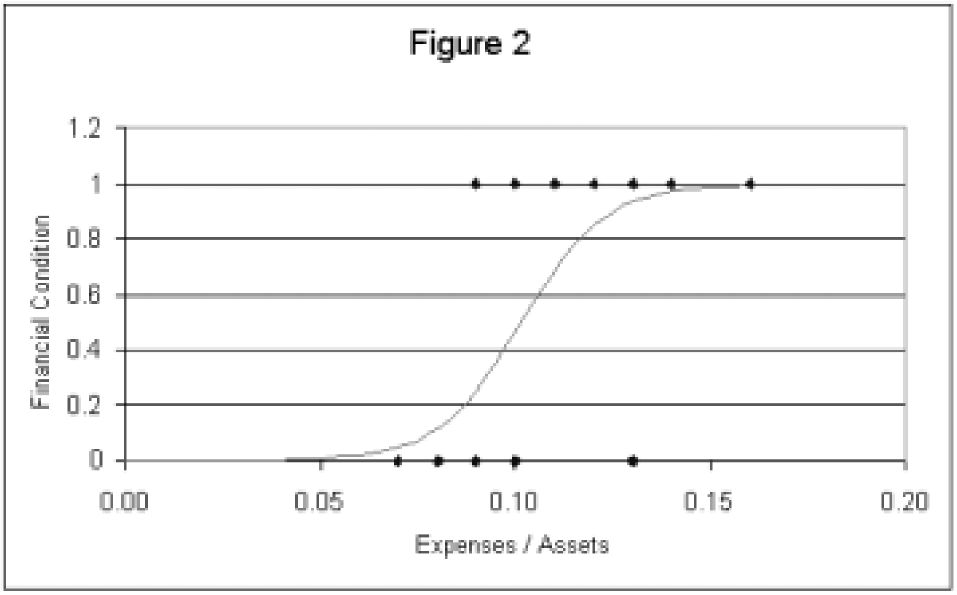
$$Prob(Y = l|x_1) = \frac{\exp(\beta_0 + \beta_2 * x_2)}{1 + \exp(\beta_0 + \beta_2 * x_2)}$$

or, equivalently,

$$\text{Odds} \ \ (Y = 1 \ \ \text{versus} \ \ Y = 0) = (\beta_0 + \beta_2 * x_2).$$

The Maximum Likelihood Estimates of the coefficients for the model are: $\beta_0 = -9.587$, $\beta_2 = 94.345$

Figure 2 displays the data points and the fitted logistic regression model.



Computation of Estimates

As illustrated in Examples 1 and 2, estimation of coefficients is usually carried out based on the principle of *maximum likelihood* which ensures good asymptotic (large sample) properties for the estimates. Under very general conditions maximum likelihood estimators are:

- *Consistent* : the probability of the estimator differing from the true value approaches zero with increasing sample size;

- *Asymptotically Efficient* : the variance is the smallest possible among consistent estimators

- *Asymptotically Normally Distributed*: This allows us to compute confidence intervals and perform statistical tests in a manner analogous to the analysis of linear multiple regression models, provided the sample size is 'large'.

Algorithms to compute the coefficient estimates and confidence intervals are iterative and less robust than algorithms for linear regression. Computed estimates are generally reliable for well-behaved datasets where the number of observations with depende nt variable values of both 0 and 1 are ‘large’; their ratio is ‘not too close’ to either zero or one; and when the number of coefficients in the logistic regression model is small relative to the sample size (say, no more than 10%). As with linear regression collinearity (strong correlation amongst the independent variables) can lead to computational difficulties. Computationally intensive algorithms have been developed recently that circumvent some of these difficulties [3].

Appendix A

Computing Maximum Likelihood Estimates and Confidence Intervals for Regression Coefficients

We denote the coefficients by the $p \times 1$ column vector β with the row element i equal to β_i, The n observed values of the dependent variable will be denoted by the $n \times 1$ column vector y with the row element j equal to y_j; and the corresponding values of the independent variable i by x_{ij} for

$$i = 1 \cdots p; j = 1 \cdots n.$$

Data : $y_j, x_{1j}, x_{2j}, \cdots, x_{pj}, \quad j = 1, 2, \cdots, n.$

Likelihood Function: The likelihood function, L, is the probability of the observed data viewed as a function of the parameters (β_{2i} in a logistic regression).

$$\begin{aligned}
&\prod_{j=1}^{n} \frac{e^{y_i(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}\cdots+\beta_p x_{pj}})}{1+e^{\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}\cdots+\beta_i x_{pj})}} \\
&= \prod_{j=1}^{n} \frac{e^{\Sigma_i y_j \beta_i x_{ij}}}{1+e^{\Sigma_i \beta_i x_{ij}}} \\
&= \frac{e^{\Sigma_i(\Sigma_j y_j x_{ij})\beta_i}}{\prod_{j=1}^{n}[1+e^{\Sigma_i \beta_i x_{ij}}]} \\
&= \frac{e^{\Sigma_i \beta_i t_i}}{\prod_{j=1}^{n}[1+e^{\Sigma_i \beta_i x_{ij}}]} \\
\text{where} \quad & t_i = \Sigma_j y_j x_{ij}
\end{aligned}$$

These are the sufficient statistics for a logistic regression model analogous to $\hat{y}$ and S in linear regression.

Loglikelihood Function: This is the logarithm of the likelihood function,

$$l = \Sigma_i \beta_i t_i - \Sigma_j \log[1 + e^{\Sigma_i \beta_i x_{ij}}].$$

We find the maximum likelihood estimates, $\hat{\beta}_i$, of β_i by maximizing the loglikelihood function for the observed values of y_j and x_{ij} in our data. Since maximizing the log of a function is equivalent to maximizing the function, we often work with the loglikelihood because it is generally less cumbersome to use for mathematical operations such as differentiation.

Since the likelihood function can be shown to be concave, we will find the global maximum of the function (if it exists) by equating the partial derivatives of the loglikelihood to zero and solving the resulting nonlinear equations for $\hat{\beta}_i$.

$$\begin{aligned}\frac{\partial l}{\partial \beta_i} &= t_i - \Sigma_j \frac{x_{ij} e^{\Sigma_i \beta_i x_{ij}}}{[1 + e^{\Sigma_i \beta_i x_{ij}}]} \\ &= t_i - \Sigma_j x_{ij} \hat{\pi}_j = 0, i = 1, 2, \cdots, p \\ \text{or} \quad & \Sigma_i x_{ij} \hat{\pi}_j = t_i\end{aligned}$$

where $\hat{\pi}_j = \frac{e^{\Sigma_i bb_i x_{ij}}}{[1+e^{\Sigma_i \beta_i x_{ij}}]} = E(Y_j)$

An intuitive way to understand these equations is to note that

$$\Sigma_j x_{ij} E(Y_j) = \Sigma_j x_{ij} y_j$$

In words, the maximum likelihood estimates are such that the expected value of the sufficient statistics are equal to their observed values.

Note : If the model includes the constant term $x_{ij} = 1$ for all j then $\Sigma_j E(Y_j) = \Sigma_j y_j$, i.e. the expected number of successes (responses of one) using MLE estimates of β_i equals the observed number of successes. The $\hat{\beta}_i$'s are consistent, asymptotically efficient and follow a multivariate Normal distribution (subject to mild regularity conditions).

Algorithm : A popular algorithm for computing $\hat{\beta}_i$ uses the Newton-Raphson method for maximizing twice differentiable functions of several variables (see Appendix B).

The Newton-Raphson method involves computing the following successive approximations to find $\hat{\beta}_i$, the likelihood function

$$\beta^{t+1} = \beta^t + [I(\beta^t)]^{-1} \nabla I(\beta^t)$$

where

$$I_{ij} = \frac{\partial^2 l}{\partial \beta_i \partial_j \beta_j}$$

- On convergence, the diagonal elements of $I(\beta^t)^{-1}$ give squared standard errors (approximate variance) for $\hat{\beta}_i$.

- Confidence intervals and hypothesis tests are based on asymptotic normal distribution of $\hat{\beta}_i$.

  The loglikelihood function is always negative and does not have a maximum when it can be made arbitrary close to zero. In that case the likelihood function can be made arbitrarily close to one and the first term of the loglikelihood function given above approaches infinity. In this situation the predicted probabilities for observations with $y_j = 0$ can be made arbitrarily close to 0 and those for $y_j = 1$ can be made arbitrarily close to 1 by choosing suitable very large absolute values of some β_i. This is the situation when we have a perfect model (at least in terms of the training data set)! This phenomenon is more likely to occur when the number of parameters is a large fraction (say $> 20\%$) of the number of observations.

Appendix B

The Newton-Raphson Method

This method finds the values of β_i that maximize a twice differentiable concave function, $g(\beta)$. If the function is not concave, it finds a local maximum. The method uses successive quadratic approximations to g based on Taylor series. It converges rapidly if the starting value, β^0, is reasonably close to the maximizing value, $\hat{\beta}$, of β.

The gradient vector ∇ and the Hessian matrix, H, as defined below, are used to update an estimate β^t to β^{t+1}.

$$\nabla g(\beta^t) = \begin{bmatrix} \vdots \\ \frac{\partial g}{\partial \beta_i} \\ \vdots \end{bmatrix}_{\beta^t} \qquad H(\beta^t) = \begin{bmatrix} & \vdots & \\ \cdots & \frac{\partial^2 g}{\partial \beta_i \partial \beta_k} & \cdots \\ & \vdots & \end{bmatrix}_{\beta^t} .$$

The Taylor series expansion around β^t gives us:

$$g(\beta) \simeq g(\beta^t) + \nabla g(\beta^t)(\beta - \beta^t) + 1/2(\beta - \beta^t)' H(\beta^t)(\beta - \beta^t)$$

Provided $H(\beta^t)$ is positive definite, the maximum of this approximation occurs when its derivative is zero.

$$\nabla g(\beta^t) - H(\beta^t)(\beta - \beta^t) = 0$$

or

$$\beta = \beta^t - [H(\beta^t)]^{-1} \nabla g(\beta^t).$$

This gives us a way to compute β^{t+1}, the next value in our iterations.

$$\beta^{t+1} = \beta^t - [H(\beta^t]^{-1} \nabla g(\beta^t).$$

To use this equation H should be non-singular. This is generally not a problem although sometimes numerical difficulties can arise due to collinearity.

Near the maximum the rate of convergence is quadratic as it can be shown that

$|\beta_i^{t+1} - \hat{\beta}_i| \leq c|\beta_i^t - \hat{\beta}_i|^2$ for some $c \geq 0$ when $\beta_{\mathrm{i}}^{\mathrm{t}}$ is near $\hat{\beta}_{\mathrm{i}}$ for all i.