MATHEMATICAL MODELING WITH LOGISTIC REGRESSION ANALYSIS IN EXCEL

William P. Fox, Ph.D.
Department of Defense Analysis
Naval Postgraduate School

Abstract

Excel does not have a logistic regression function. We must use the solver to numerically solve for the coefficient that maximizes the likelihood function. We need to able to create a table of results to determine how well our model works. We also want to illustrate a potential problem that can be found with any solver. We present formulas and a methodology to perform logistic regression in Excel and get the outputs required to build tabular results similar to other statistical packages.

Keywords: Excel, logistics regression, formulas. Maximize, likelihood function, ln likelihood function.

Introduction

In data analysis, **logistic regression** (sometimes called the **logistic model** or **logit model**) is a type of regression analysis used for predicting the outcome of a binary dependent variable (a variable which can take only two possible outcomes, e.g. "yes" vs. "no" or "success" vs. "failure") based on **one or more** predictor variables. Logistic regression attempts to model the probability of a "yes/success" outcome using a linear function of the predictors. Specifically, the log-odds of success (the logit of the probability) is fit to the predictors using linear regression. Logistic regression is one type of discrete choice model, which in general predict categorical dependent variables — either binary or multi-way.

Like other forms of regression analysis, logistic regression makes use of one or more predictor variables that may be either continuous or categorical. Also, like other linear regression models, the expected value (average value) of the response variable is fit to the predictors — the expected value of a Bernoulli distribution is simply the probability of success. Unlike ordinary linear regression, however, logistic regression is used for predicting binary outcomes (Bernoulli trials) rather than continuous outcomes, and models a transformation of the expected value as a linear function of the predictors, rather than the expected value itself.

For example, logistic regression might be used to predict whether a patient has a given disease (e.g. diabetes), based on observed characteristics of the patient (age, gender, body mass index, results of various blood tests, etc.). Another example might be to predict whether a voter will vote Democratic or Republican, based on age, income, gender, race, state of residence, votes in previous elections, etc. Logistic regression is used extensively in numerous disciplines: the medical and social sciences fields, natural language processing, marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription, etc. Yes, even in the military, logistic regression has utility as we will see.

The model for just one predictor is

$$Y_i = \frac{B_0}{1+B_1 e^{B_2 X_i}} + \varepsilon_i$$

where the error terms are independent and identically distributed (*iid*) as normal random variables with constant variance.

For more than one predictor we use the model

$$Y_{i=}\frac{e(b_0+\Sigma b_i x_i)}{1+e(b_0+\Sigma b_i x_i)}$$

What is Logistic Regression?

Logistic Regression calculates the probability of the event occurring, such as the purchase of a product. In general, the object being predicted in a regression equation is represented by the dependent variable or output variable and is usually labeled as the Y variable in the Regression equation. In the case of Logistic Regression, this "Y" is binary. In other words, the output or dependent variable can only take the values of 1 or 0. The predicted event either occurs or it doesn't occur – your prospect either will buy or won't buy. Occasionally this type of output variable is also referred to as a Dummy Dependent Variable.

Output Desired

We assume we would like to obtain as much output as possible, but at a minimum we want:

Estimates of the coefficients, their standard errors, t statistics, P-values, and some analysis of fit between the full model and a not-full model that includes -2 ln likelihood and chi-squared tests.*

An Example of Logistic Regression

To simplify the analysis in Excel, we create a maximum ln likelihood function for the logit expression:

$$ln\ L(B_i)= \Sigma Yi(B0+\Sigma B_i X_i) -\Sigma ln(1+exp(B0+\Sigma B_i X_i))$$

I know this looks intimidating but it is not. We can build the functions and optimize in EXCEL using this function.

Example 1. We have the following data where the response, Y, is a binomial from Bernoulli trials—like yes or no. In this case it is a success, 1, or a failure, 0.

Item	Status	Number	Difference
1	1	4	19.2
2	1	2	24.1
3	0	4	-7.1
4	1	3	3.9
5	0	9	4.5
6	0	6	10.6
7	0	2	-3
8	0	11	16.2
9	1	6	72.8
10	0	7	28.7
11	1	3	11.5
12	1	2	56.3
13	0	5	-0.5
14	0	3	-1.3
15	0	3	12.9
16	0	8	34.1
17	0	10	6.6
18	1	5	-2.5
19	0	13	24.2
20	0	7	2.3
21	1	3	36.9
22	0	4	-11.7
23	1	2	2.1
24	1	3	10.4
25	0	2	9.1
26	0	5	2
27	0	6	12.6
28	1	5	18
29	0	3	1.5
30	1	4	27.3
31	0	10	-8.4

The model we want is

$$Y_i=\frac{e^{(b_o+b_1 number+b_2 difference)}}{1+e^{(b_o+b_1 number+b_2 difference)}}$$

This is how we proceed:

(1) We entered the data.
(2) Create heading and initial values for our model's coefficients: B_0, B_1, and B_2. Usually we set them at 0.
(3) Create the functions we need (I do this in two parts). Column P1 uses

Yi(B0+B1*number+B2*difference)*
and Column P uses
*ln(1+exp(B0+B1*number+B2*difference)).*

(4) *Sum columns P1 and P2.*
(5) *In an unused cell take the difference of P1-P2—this is the objective function.*
(6) *Open the Solver and Maximize this cell containing P1-P2, by changing cells with B0,B1, and B2. Insure to uncheck the non-negativity box.*
(7) *Solve.*
(8) *Obtain your model and use it as needed.*
(9) *Repeat steps 3-8 for the model with intercept only.*

We have the data entered in to three columns.

Next we create columns for **Y*X'B** and *ln(1+exp(**X'B**))* using initial values for b0, b1, and b2.

We sum these two columns separately and in another cell we take the difference in the sums.

This is our objective function that we maximize by changing the cells for b0,b1, and b2. By doing so we get the following results .

Bo	*B1*	*B2*
1.421207	-0.75534	0.112205

P1	Yi(B0+B1Number+B2*difference)	P2	LN(1+exp(Bo+B1*number+B2*difference)
0.554174			1.008141
2.614663			2.685301
0			0.087101
-0.40722			0.510124
0			0.007629
0			0.136619
0			0.502626
0			0.006264
5.057677			5.064017
0			0.421461
0.445538			0.940526
6.227665			6.229637
0			0.085876
0			0.315775
0			1.039183
0			0.372547
0			0.004544
-2.63602			0.069196
0			0.003398
0			0.026742
3.295545			3.331923
0			0.052891
2.390253			2.477904
0.322112			0.867117
0			1.263713
0			0.112174
0			0.16824
-0.33581			0.53927
0			0.411041
1.463034			1.671294
0			0.000846
18.99161			30.41312
-11.4215			

We obtain our model,

*Y=exp(1.42120.7553*number+0.1122*difference)/ (1+exp(1.42120.7553*number+0.1122*difference))*

A plot of this model is shown in Figure 1.

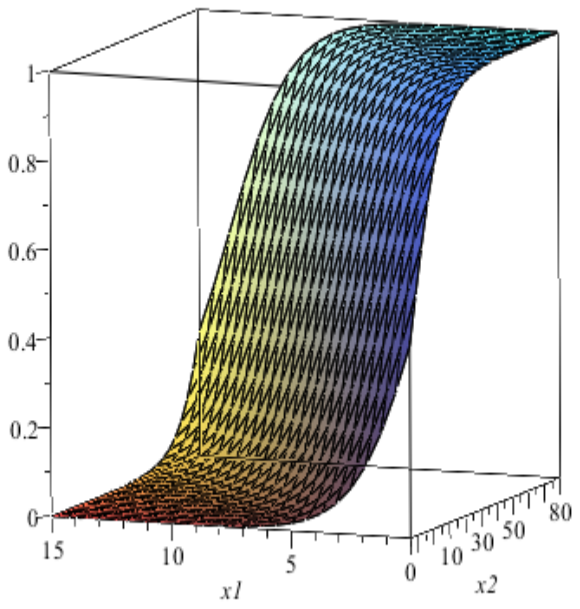

Figure 1. Plot of the Logistic Function (in this case a 3D CDF).

Before we accept this model, we require a minimum of a few diagnostics. We want to (1) examine the significance of each estimated coefficient {b0,b1,b2} and (2) compare this full model to an intercept only model and one term model to measure the chi-square differences.

We start with the estimates of our full model's coefficients {b0=1.421207, b1=-0.75534, and b2=0.112205}. We need the following (a) estimates of the standard errors of these estimates, (b) t* which equals the *estimates/se*, and (c) P-value for t*.

We know that the estimates for the Variance-Covariance matrix are the inverse of the Hessian matrix evaluated at the estimates of {b0, b1, and b2}. In a logistics equation the number of terms in the regression model affects the Hessian matrix. To obtain all this we will need the Hessian matrix, *H(X)*, so that we can find the inverse, $H(X)^{-1}$, and then - $H(X)^{-1}$ that is the variance-covariance matrix when evaluated at our final coefficient estimates. The main diagonal of this matrix are our Variances for each coefficient. If we take the square root of these coefficients then we get the *se* of our coefficients.

$$\left[\left[-\left(\sum_{i=1}^{n}\left(\frac{e^{b0+b1\,x(i)}}{1+e^{b0+b1\,x(i)}}-\frac{\left(e^{b0+b1\,x(i)}\right)^2}{\left(1+e^{b0+b1\,x(i)}\right)^2}\right)\right),-\left(\sum_{i=1}^{n}\left(\frac{x(i)\,e^{b0+b1\,x(i)}}{1+e^{b0+b1\,x(i)}}-\frac{\left(e^{b0+b1\,x(i)}\right)^2 x(i)}{\left(1+e^{b0+b1\,x(i)}\right)^2}\right)\right)\right],\left[-\left(\sum_{i=1}^{n}\left(\frac{x(i)\,e^{b0+b1\,x(i)}}{1+e^{b0+b1\,x(i)}}-\frac{\left(e^{b0+b1\,x(i)}\right)^2 x(i)}{\left(1+e^{b0+b1\,x(i)}\right)^2}\right)\right),-\left(\sum_{i=1}^{n}\left(\frac{x(i)^2\,e^{b0+b1\,x(i)}}{1+e^{b0+b1\,x(i)}}-\frac{x(i)^2\left(e^{b0+b1\,x(i)}\right)^2}{\left(1+e^{b0+b1\,x(i)}\right)^2}\right)\right)\right]\right]$$

We would like to use pattern recognition and a simplification step to better see what is happening here and let

$$\pi = exp(b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n).$$

Let's let $P = -\sum_{i=1}^{n}\left(\frac{\pi}{1+\pi}\right) - \left(\frac{\pi^2}{(1+\pi)^2}\right)$. Then we can more easily write the Hessian matrix for n terms and its inverse as follows:

$$\begin{bmatrix} \sum P & \sum x_1 P & \sum x_2 P & \ldots & \sum x_n P \\ \sum x_1 P & \sum x_1^2 P & \sum x_1 x_2 P & \ldots & \sum x_1 x_n P \\ \sum x_2 P & \sum x_1 x_2 P & \sum x_2^2 P & \ldots & \sum x_2 x_n P \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_n P & \sum x_1 x_n P & \sum x_2 x_n P & & \sum x_n^2 P \end{bmatrix}$$

$$\begin{bmatrix} \sum P & \sum x_1 P & \sum x_2 P & \ldots & \sum x_n P \\ \sum x_1 P & \sum x_1^2 P & \sum x_1 x_2 P & \ldots & \sum x_1 x_n P \\ \sum x_2 P & \sum x_1 x_2 P & \sum x_2^2 P & \ldots & \sum x_2 x_n P \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_n P & \sum x_1 x_n P & \sum x_2 x_n P & & \sum x_n^2 P \end{bmatrix}^{-1}$$

We take the square root of the entries on the main diagonal as our estimates of the *se* for $\{b_0, b_1, \ldots, b_n\}$.

In our example we compute H and H^{-1}. We compute H using the sums of the columns in the matrix H. To obtain H-1, we use =MINVERSE command in Excel.

H

3.711818969	15.00133	44.21611
15.00132821	72.48339	218.9457
44.21611161	218.9457	1041.944

H^-1

1.655778953	-0.35711	0.004774
-0.35710513	0.114788	-0.00897
0.004774266	-0.00897	0.002641

We take the square root of the main diagonal entries {1.655778953, 0.114788, 0.002641} as our standard error, *se*, estimates for b0, b1, and b2, respectively. We find that our estimates are { 1.286770746, 0.338803, 0.051393}.

We can enter and fill in the rest of our table.

Analysis of regression coefficients.

Coefficient	Estimate from the Solver	*Se from the square root of the V-C Matrix*	Z-statistic= *estimate/se*	*P-Value from P(Z>|Z-statistic|)*
b_0	1.421207	1.286770746	1.1044	0.2694
b_1	-0.75534	0.338803	-2.2296	0.02577
b2	0.112205	0.051393	2.162	0.0306

We see from the results in the table that the coefficients for b_1 and b_2 are significant at α=0.05.

Let's calculate the deviances for our model. We define the ***deviances*** as

$$dev_i = \pm[-2[Y_i \ln(\pi_{1i}) + (1 - Y_{i)} \ln(1 - \pi_{1i})]]^{\frac{1}{2}}$$

where the sign is positive when $Y_i \geq \pi_{1i}$ and negative when $Y_i < \pi_{1i}$ and we define $\pi_1$as $(1+\pi)^{-1}$ with π as we defined earlier.

Analysis of Deviations

Model	*ln* likelihood	Deviance	*df*	Chi-square	P-value
Full Model	-11.4215	22.842	3		
Constant Model	-20.6904	41.3808	1		
Difference	-9.26887	18.53774	2	18.5377	$9.43x10^{-5}$

We find the difference is significant at α=0.05 so we choose the full model over the constant model.

Odds-Ratios

Interpretation of *B* parameters in the logistic model.

$\pi^*=B_0+B_1x_1+\ldots+B_nx_n$

where

π^*

π^*

B_1=Change in log-odds π^* for every 1 unit increase in x_1 holding all other x's fixed.

$e^{Bi}-1$ = Percentage change in odds ratio $\pi/(1-\pi)$for every 1 unit increase in x_1 holding all other x's fixed.

So, $B_1=-0.7553$, $e^{B1}=0.47$, $e^{B1}-1=-0.53$. For each unit of x_1, we estimate the odds of a fixed contract to decrease by *53%* holding x_2 fixed.

For $B_2=0.1122$,$e^{B2}=1.12$,$e^{B2}-1=0.12$. For each unit of x_2, we estimate the odds of a fixed contract to increase by *12%* holding x_1 fixed.

Conclusion

We have presented a "how to" approach to logistic regression in Excel. We have given formulas and suggested how to put them into a table so that good analysis can be made.

References

1. Fox, William P. (2012) Mathematical Modeling with Maple. Cengage Publishers. Boston, MA.

2. Giordano, F. W. Fox, S. Horton, & M. Weir. (2008) A First Course in Mathematical Modeling. Cengage Publishing. Belmont, CA.

3. Fox, William P. (2011). Using the EXCEL Solver for Nonlinear Regression, *Computers in Education Journal (COED*). October-December, **2**(4), pp. 77-86.

4. Fox, William P. (2012). Issues and Importance of 'Good" Starting Points for Nonlinear regression for Mathematical Modeling with Maple: Basic Model Fitting to Make Predictions with Oscillating Data. *Journal of Computers in Mathematics and Science Teaching. **31**(1), pp. 1-16.*

5. Mendenhall, W. & T. Sincich. (1996). A Second Course in Statistics Regression Analysis, 5th ed. Prentice Hall, Upper saddle River, NJ. pp. 476-485.

6. Montgomery, D., E. Peck, & G.Vinning. (2006). Introduction to Linear Regression Analysis, 4th ed. John Wiley and Sons. Hoboken, NJ. pp. 428-448.

Biographical Information

Dr. William P. Fox is a professor in the Department of Defense Analysis at the Naval Postgraduate School. He received his BS degree from the United States Military Academy at West Point, New York, his MS from the Naval Postgraduate School, and his Ph.D. from Clemson University. Previously, he has taught at the United States Military Academy and Francis Marion University where he was the chair of the mathematics department for eight years. He has many publications including books, chapters, journal articles, conference presentations, and workshops. He directs several math modeling contests through COMAP. His interests include applied mathematics, optimization (linear and nonlinear), mathematical modeling, statistical models for medical research, and computer simulations. He is Vice-President of the Military Application Society in INFORMS.