The Logit Regression Model

Why regression does not work

With binary or categorical dependent variables, standard regression analysis is not appropriate. In the examples that follow, this means that we are using a binary input column to, for example, test whether we can predict examination success or failure: 1 means success and 0 means failure.

Example

- binary dependent variable y coded to be 0 for non purchases and 1 for purchases
- X is a continuous metric variable

$$y = \alpha + \beta x + \varepsilon \quad \text{with}$$
$$y = \begin{cases} 0 & \text{for non purchases} \\ 1 & \text{for purchases} \end{cases}$$

Problems

- After least square estimation predictions of y using the value of x would produce many other values than zero and one including values below zero and values above one
- Different coding for the binary dependent variable (eg 1 and 2, or 0 and 10) would lead to very different estimates for the α and β coefficients which makes the interpretation of the regression parameters difficult
- The above model does not meet the assumptions of the regression model since multivariate normality of the dependent variable for any value of the explanatory variables is broken

Binary Choice Model

- Y can assume the discrete values of 0 or 1
- To model y as a function of x, one can exploit a latent variable
- y assumes either value 0 or 1 depending on the threshold value
- δ of a metric and continuous Latent variable z

The regression model is rewritten as

y, = 0 if z; < δ
y. = 1 if z; > δ

- The dependent variable y is 1 when a latent continuous variable z is above the threshold δ and 0 otherwise
- The model is completed by a regression equation linking the latent variable to the explanatory variable

$$z_i = \alpha + \beta x_i + \varepsilon_i$$

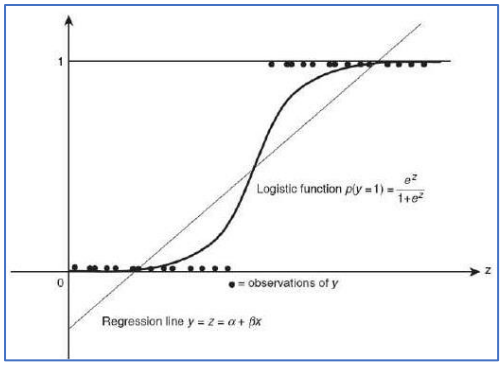
The logistic transformation

- Probabilities that y = 1 (on the vertical axis) concentrate around 0 for values of x below a

certain threshold, then go quickly towards 1 when x is above the threshold.

- The function fits well with the need for approximating the probabilities of a binary outcome as a function of the explanatory variable.
- The logistic transformation of y into z is obtained by applying the logit link function to the expected value of y

.



Logistic regression

- The logit transformation is the link function for logistic regression
- The logit transformation is the log of the odds that y = 1 relative to y = 0
- The logit link allows to transform the binary variable y into a continuous variable z
- The final equation is a regression model with a continuous variable on the left hand side
- The only difference from the standard regression model is that the distribution of the error is not normal but logistic.
- Estimation of α and β can be obtained by maximum likelihood which works with any known probability distribution of the errors and returns the maximum likelihood estimates (the most probable values for the parameters)

Types of discrete choice models

- Logistic regression: at least one of the explanatory variables is metric and continuous
- Logit model: all of the variables on the right hand side are non metric (binary or categorical)
- In a logit model with a categorical or binary x variable, the coefficient β is mathematically related to the odds ratio (with respect to the baseline category of x) of having a positive outcome
  - For example, if the dependent variable is one when the consumer buys a specific brand and x measures whether the consumer has kids or not, one can compute with e^{β} the odds ratio of buying the brand for consumers with kids as compared to consumers without kids.

Probit model

- The Probit model is also applied to binary dependent variables but with different assumptions on the link function and the error distribution
- The link function (called probit) is the inverse of the standard normal cumulative distribution function
- This link function guarantees that the distribution of the model which is finally estimated is still normal
- The choice between the probit and the logit distribution depends on the type of

dependent variable

- if the dependent variable can be reasonably assumed to be a proxy for a true underlying variable which is normally distributed then the probit model should be chosen
- if the dependent variable is considered to be a truly qualitative and binomial character then logit modelling should be preferred
- generally the two models lead to very similar results, unless cases are concentrated to the tails of the distributions in which case the logit link function should be chosen

The Excel Examples

The file *logistic_regression.xlsx* that accompanies this discussion illustrates four logit models that help us to understand and apply the method in Excel. Of course, this is Excel so we have to program the technique, as opposed to, eg, R or SPSS or any other dedicated statistical software package.

In example 1 we would like to know how profitable it is, in terms of examination success, to spend more or less time preparing for an exam! The screenshot below shows the data and the intermediate columns that we must set up to find the logit model.

In the discussion above we used alpha and beta in our equation, in my Excel file, I have used b_0 and b_1 instead and they are entered in cells B5 and B6 respectively.

Note that, we set b_0 and b_1 to 0.001 when we start this exercise because we have no idea what they should be. Using 0.001 or any other value will allow us to develop the model and see how it works as we prepare it.

The final step in modelling this example is to use SOLVER to find the maximum possible values of b_0 and b_1

I have left the final answer in my screenshot, so you can see 0.001 for b_0 has become -4.0777 and b_1 has become 1.5046

	A	B	C	D	E	F
1	**Logistic Regression**					
2	https://en.wikipedia.org/wiki/Logistic_regression					
3						
4	set all of these to 0.001 initially					
5	b0	(4.0777)				
6	b1	1.5046				
7						
8					Total	(8.0299)
9	**Pass/Fail**	**Hours of Study**	**Logit**	**eLogit**	**Probability**	**Log Likelihood**
10	0	0.5	(3.3254)	0.0360	0.0347	(0.0353)
11	0	0.75	(2.9492)	0.0524	0.0498	(0.0511)
12	0	1	(2.5731)	0.0763	0.0709	(0.0735)
13	0	1.25	(2.1969)	0.1111	0.1000	(0.1054)
14	0	1.5	(1.8207)	0.1619	0.1393	(0.1501)
15	0	1.75	(1.4446)	0.2358	0.1908	(0.2118)
16	1	1.75	(1.4446)	0.2358	0.1908	(1.6563)
17	0	2	(1.0684)	0.3436	0.2557	(0.2953)
18	1	2.25	(0.6923)	0.5004	0.3335	(1.0980)
19	0	2.5	(0.3161)	0.7290	0.4216	(0.5475)
20	1	2.75	0.0601	1.0619	0.5150	(0.6636)
21	0	3	0.4362	1.5468	0.6074	(0.9349)
22	1	3.25	0.8124	2.2533	0.6926	(0.3673)
23	0	3.5	1.1885	3.2823	0.7665	(1.4545)
24	1	4	1.9409	6.9648	0.8744	(0.1342)
25	1	4.25	2.3170	10.1454	0.9103	(0.0940)
26	1	4.5	2.6932	14.7786	0.9366	(0.0655)
27	1	4.75	3.0693	21.5278	0.9556	(0.0454)
28	1	5	3.4455	31.3591	0.9691	(0.0314)
29	1	5.5	4.1978	66.5415	0.9852	(0.0149)

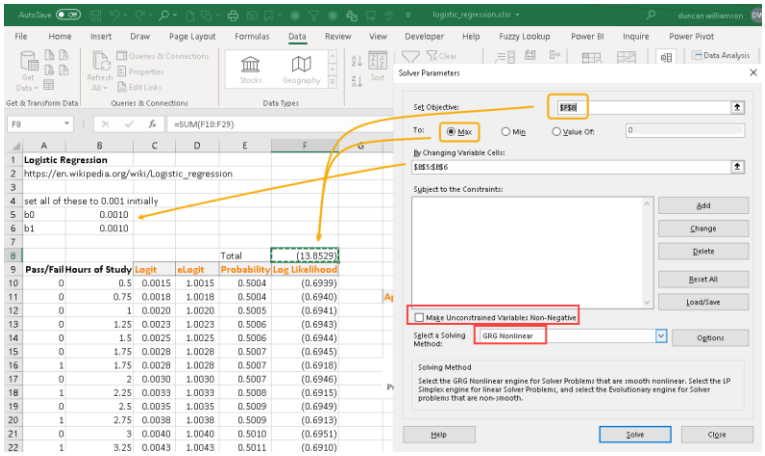
The formulas you need are:

Logit = C10=B$5+B$6*B10
eLogit = D10 =EXP(C10)
Probability= E10 =D10/(1+D10)
Log Likelihood = F10 =A10*LN(E10)+(1-A10)*LN(1-E10)

Fill down, in this case to row 29 in all columns.

Using SOLVER to find b_0 and b_1

We initially set b_0 and b_1 to 0.001, as discussed above. Before we find our final answer, we need to use SOLVER to find their optimal values. Here is a screenshot of my SOLVER workings:



Firstly, I had reset b_0 and b_1 to 0.001
Secondly, the Total in cell F8 is the sum of the Log Likelihood column

Data … Analyse … SOLVER

Set the Objective to cell B8 and the Changing Variable cells to B5:B6
Note that we UNcheck the Non Negativity box
We are using the GRG Nonlinear function since Logit regression is a non linear model.

OK

As I mentioned earlier
B0 changes from 0.001 to -4.0777
B1 changes from 0.001 to 1.5046

The output is, using 5 hours of study:

Application			
	Hours of Study	5	
		Log Odds	3.4455
		Odds	31.3591
		Pr	0.9691

And the formulas you need here are

Hours of Study … you choose … I put 5 for demonstration

Log Odds = K13 =B5+B6*J12 … note we are using b_0 and b_1 here
Odds = K14 =EXP(B6*J12+B5) … again, b_0 and b_1
Probability = K15 =1/(1+EXP(-(B6*J12+B5))) … and again, b_0 and b_1

Interpretation

The interpretation of the results means interpreting the Pr result. In the screenshot above, we tested for 5 hours of study and whether that would give us a good result on our test or examination and, good news, based on the data provided, 5 hours gives us a 96.91% chance of success!

I amended my template so that I could get more information from the model:

Applicant	19			
Actual Hours	5.00			
		Log Odds	3.4455	
		Odds	31.3591	
		Pr	0.9691	
		Admission Prediction		1
		Admission Actual		1

Now I have tied my question and answer to one of the actual examinees by adding an Examinee number column in column A and then the above output. This shows that candidate 19 studied for 5 hours and he passed the examination. The Logit model says he should have passed, too!

The formulas I used are here:

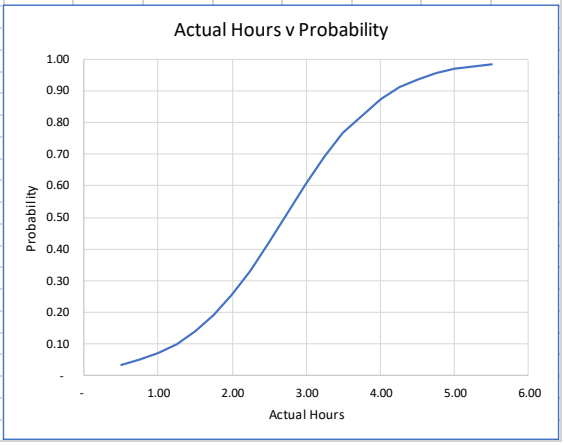
Applicant	1				
Actual Hours	0.50				=VLOOKUP(J17,A9:C29,3,0)
		Log Odds	(3.3254)		=B5+B6*J18
		Odds	0.0360		=EXP(B6*J18+B5)
		Pr	0.0347		=1/(1+EXP(-(B6*J18+B5)))
		Admission Prediction		0	=IF(L21<0.5,0,1)
		Admission Actual		0	=VLOOKUP(J17,A9:C29,2,0)

Overall Accuracy of the Logit Model

Let's test the model to see how many predictions it got right. I ran the model using all 20 candidates' actual hours of study and in all but 4 cases, the model predicted the right outcome: that's an 80% predictive success rate. Candidates 7, 9, 12 and 14 gave the wrong prediction.

Overall, the outcomes in a table and graph look like this:

Actual Hours	Probability
0.50	0.03
0.75	0.05
1.00	0.07
1.25	0.10
1.50	0.14
1.75	0.19
1.75	0.19
2.00	0.26
2.25	0.33
2.50	0.42
2.75	0.52
3.00	0.61
3.25	0.69
3.50	0.77
4.00	0.87
4.25	0.91
4.50	0.94
4.75	0.96
5.00	0.97
5.50	0.99



Conclusions

Learning Logit regression is a useful technique to learn and by using the template I have created here, you can copy and paste your data in my Example 1 template and check the answers. Please note, you would need to change the template if you used more or less than 20 input rows and if you used more variables. See my examples 2, 3 and 4, though, which do use different numbers of input rows and columns.

For reference:

Example 2 has 12 input rows, 4 input columns and 4 output columns
Example 3 has 400 input rows, 4 input columns and 4 output columns
Example 4 has 20 input rows, 2 input columns and 4 output columns

Duncan Williamson
27th October 2020