WHAT IS LOGISTIC REGRESSION

Joseph M. Hilbe

13 March, 2013

1: The Logistic Model

Logistic regression is a single parameter discrete response regression model where the response is either binary (0,1), or is partitioned into a numerator (*y*) and denominator (*m*), with the denominator being the number of observations having the same pattern of covariates and the numerator being a count of the number of observations where *y*==1 for each covariate pattern. The response variable is then a combined y/m. The latter parameterization of logistic regression is typically referred to as *grouped logistic regression*, and was the first way that the logistic model was used. I address binary logistic regression in the remainder of this paper.

The fitted or predicted value for a binary response logistic model is a probability; ie. the probability that *y*==1 (some software algorithms use *y*==0 rather than *y*==1), where 1 is regarded as the binary "success" and 0 as non-success or "failure". For example, if the binary response or dependent variable *y* is "patient died while in hospital", then *y*==1 could be regarded as "yes" (they did die while in the hospital) and *y*==0 as "no" (patient did not die while in hospital).

When a logistic model is estimated using GLM or Generalized Linear Models (SAS=Prov GENMOD, Stata=glm, R=glm(), SPSS=GENLIN), the predicted value is usually referred to as *mu,* or μ (the Greek letter). The predicted values may be obtained by applying the GLM inverse link function to the linear predictor, which is SUM(X'*Beta) or the sum of the products of the predictor values and predictor coefficients, Σ*x*'β. We can refer to the linear predictor as eta (η) or *x*'β, Over all of the observations and predictors in a model, the regression may be symbolized as,

$$\sum_{i=1}^{n} \beta_0 + x'_1\beta_1 + x'_2\beta_2 + \cdots + x'_n\beta_n$$

The logistic inverse link is 1/(1+exp(-XB)) or exp(XB)/(1+exp(XB)), which converts the linear predictor to the fitted value, μ. Therefore,

$$\mu = \frac{1}{1+\exp\,(-x\beta)} \;\; or \;\; \frac{exp(x\beta)}{1+exp(x\beta)}$$

for each observation in the model.

The logistic link function is the basis of understanding, and calculating, the above inverse link function. First, we understand that μ is the probability that the response variable, *y*, is 1. Hence,

$$\mu = Pr(y == 1)$$

Next, the definition of odds is the probability of success divided by the probability of failure. In terms of μ, the odds is μ/(1-μ). Many times this relationship is displayed as p/(1-p) where p is the probability of success. For the binary (1,0) response logistic model, μ=p.

A key to understanding the binary logistic model, and what allows logistic regression to be estimated as a GLM (generalized linear model) is realize that if μ=1/(1+exp(-*x*β)), then,

$$x'\beta = log\left(\frac{\mu}{1-\mu}\right)$$

which is the natural log of the odds of μ --- better known as the **logit**. It is the logistic link function, which relates the linear predictor in terms of the fitted value. For the logistic model, the link relates the linear predictor of the model to the probability that *y*==1. The logit link function gave the name to the model as logit regression,, which is how the model is generally referred to in econometrics. Most of areas for study refer to the model as logistic regression. The logistic regression model can then be symbolized as:

$$log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{i=1}^{n_i} \beta_0 + x'_{i,1}\beta_1 + x'_{i,2}\beta_2 + \cdots + x'_{i,n}\beta_n$$

You will many times see that the subscripts of the predictors are simply given are i, but technically each row can have a different value of each predictor in the model. The parameter value though is the same across observations for each predictor.

2: Estimation

Estimation of the logistic model is nearly always performed using either a full maximum likelihood algorithm, or by modeling it as a generalized linear model (GLM). Note that generalized linear models differs from the general linear model, which is based only on the normal or Gaussian distribution. GLMs include nonlinear models that can be linearized using the link function, although the normal model is also a GLM. A restriction for GLM models exits such that GLM models are all based on probability distributions belonging to the single parameter exponential family of distributions.

The commands or functions for GLM and full maximum likelihood estimation of the logistic model for four leading software applications is given below.

	GLM	MLE
Stata	glm	logit or logistic
R	glm	ml_glm (in *msme* package)
SAS	Genmod	Proc Logistic
SPSS	Genlin	Logistic

The results of using a GLM or MLE function for estimation of a logistic model are the same.
It must be understood though that GLM is in fact a type of MLE, but is a simplified version employing an iteratively reweighted least squares (IRLS) algorithm. A full explanation of methods of estimation can be found in Hilbe (2009), *Logistic Regression Models*, or in Hilbe &

Robinson (2013), *Methods of Statistical Model Estimation*, both published by Chapman & Hall/CRC.

3. Odds Ratios

Another very nice feature of logistic regression, which relates to its logit link function, is the fact that the exponentiation of a model coefficient gives the odds ratio for the predictor. The ratio is the odds of X==1 compared to X==0 for a binary predictor, or X=level of interest compared to X=reference level for categorical predictors, and X = x+1 to X=x for continuous predictors. For example, suppose that we have a binary logistic model with *y* as *died* and a single predictor, *gender*, where 1=*female* and 0=*male*, *died*==0 indicates that the patient did not die (within some specified period). Also suppose that the exponentiated coefficient on gender is 2.0. If gender significantly contributes to the model, ie tha the p-value is less than 0.05, and the model is well fitted, then we may assert that the odds of a patient dying in the hospital is twice a great for females as for males.

Suppose that we have a response , or dependent, variable of a*dmit*, with 1=admitted to college of first choice; 0=no admitted to college of first choice. Also suppose that we have a single categorical predictor, *gpa,* with 5 levels,

gpa	*OR*
1=below 2.00	0.0
2=2.00-2.49	1.5
3=2.50-2.99	2.0
4=3.00-3.49	5.0
5=3.50-4.00	9.0

To model *admit* on *gpa*, one of the levels must be selected as the reference level; ie, the level which as a binary predictor would be x==0. Stata and R by default select the first level as the reference; SAS defaults to the last level. However, you should select the reference level which makes most sense for the data being modeled. We shall use the first level as the reference, that the student had a cumulative high school *gpa* below 2.0. The estimated odds ratios, which are calculated as the exponentiation of the model coefficients, are displayed to the right of the level in the table above. Based on the model, a student with a cumulative GPA of 3.2 has 9 times the odds of getting into the first college of choice than does a student (in the model) with a cumulative GPA of under 2.0. A student with a GPA of 3.6 has 9 times the odds of getting admitted compared to a student with a sub 2.0 GPA.

If we wish to know the difference in odds between the 5th and 4th levels, it would be the same as having the 4th level as the reference, and checking the odds ratio of the 5th level. It will be the same as the 5th level divided by the 4th in the original model, or 9/5 = 1.8. That is, a student with a GPA of 3.5 or more has 1.8 times the odds of being admitted to their first choice college as does a student with a GPA between 3.00 and 3.49.

For a continuous predictor, eg. age, if the odds ratio is 0.03, the odds y==1 is 0.03 times greater for a person age 40 compared to a person age 39. The same odds ratio exists for any two consecutive values of a the predictor. If there are other predictors in the model, their values are held at a constant. The default is that other predictors are kept at their mean value.

4. Conclusion

Logistic regression models are likely the most used regression model in research after the basic normal or Gaussian model --- linear regression. I believe that all researchers should have a solid foundation in the logistic-based modeling, which extends to grouped logistic models, proportional odds models, cumulative ratio models, multinomial models, and a variety of other logistic-based models. What characterizes all members of the class of logistic models is that the predicted value is a probability, and the exponentiated coefficients are odds ratios.