Logistic Regression: The Standard Method of Analysis in Medical Research

Sloan Rush

April 23, 2001

Abstract

Regression methods are essential to any data analysis which attempts to describe the relationship between a response variable and any number of predictor variables. Frequently, situations involving discrete variables arise. For example, in a medical setting, an outcome might be presence/absence of disease. Logistic regression analysis extends the techniques of multiple regression analysis to research situations in which the outcome variable is categorical, that is, taking on two or more possible values. In this paper, the risk factors for a disease of the eye (retinopathy of prematurity) are identified using logistic regression analysis.

1 Introduction

In clinical situations, the status of a patient is assessed by the presence or absence of a disease. There are many factors to consider which may or may not correlate with the incidence of the disease. There have been numerous retrospective medical research studies published each year that review past medical records and charts of former patients to help determine some of the risk factors (or causing agents) of diseases that are of interest. Finding the risk factors and the potential risk factors can help prevent the development of the disease. All of the diseases and nearly all of the risk factors considered are categorical variables (variables taking on two or more possible values). Hosmer and Lemeshow, two prominent statisticians, state that "the logistic regression model has become the standard method of analysis in this situation."[1]

Like any other model building technique, the goal of the logistic regression analysis is "to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables."[2]

[1]Hosmer, D. W.; Lemeshow, S. (1989). Applied Logistic Regression, New York: Wiley.
[2]Hosmer, op. cit.

This statement motivates the purpose of this project: to identify risk factors for a specific disease of the eye using the statistical tool of logistic regression analysis.

2 The Disease

The disease of the eye that is studied in this paper is retinopathy of prematurity (ROP). ROP is a disease of the eye affecting primarily premature infants and is the leading cause of blindness in newborn infants. Low birth weight and low gestational age neonates are most susceptible to ROP. It is a vasoproliferative disorder of immature blood vessels in the developing eye of newborn infants. That is, the blood vessels in the eye stop growing outward in the retina and form a circular ridge of conglomerated blood vessels. A clearly defined ridge is considered Stage I ROP. As the disease progresses into Stage II and Stage III, the ridge proliferates and pushes the retina outward. Untreated Stage III ROP, can eventually result in a retinal detachment (blindness), which is Stage IV ROP.

3 The Data

The data used in this project was obtained from a medical retrospective research project, done by the author, at Northwest Texas Hospital in Amarillo, Texas. Since the project is retrospective, the data was collected from the medical records storage room at the hospital. The project was approved by the Institutional Review Board (IRB) of Texas Tech School of Medicine. The research project analyzed the incidence of the disease, retinopathy of prematurity (ROP).

The study included 300 subjects[3] from which data of 29 different risk factors[4] was collected. (The 300 subjects included in the study were all those admitted to the hospital from January 1995 through May 2000 fitting certain screening criteria). All 29 of the proposed risk factors are discrete or categorical variables, some dichotomous (yes/no) and others containing up to 5 possibilities. For example, one of the risk factors looked at was whether or not the neonate had necrotizing enterocolitis (NEC), which is classified into 4 grades or stages of development. If they did, the worst grade of NEC attained during the course of their hospital stay was recorded. For this particular predictor variable, we see that there are 5 possible outcomes: no NEC, Grade I, Grade II, Grade III, or Grade IV. Given the categorical nature of these predictor variables, we see that logistic regression analysis is suitable to model the data.

[3]Note: Of the 300 eligible subjects, only 252 will be included in the statistical model since 48 of the subjects expired.

[4]See Appendix for a brief description of the risk factors used in this study.

4 The Model

In setting up the logistic regression model, we must first establish the fundamental model for any multiple regression analysis. It is assumed that the outcome variable is a linear combination of a set of predictors. For outcome variable Y, and a set of n predictor variables, $X_1, X_2, ..., X_n$, we have the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon = \beta_0 + \sum_{j=1}^{n} \beta_j X_j + \varepsilon$$

where β_0 is the expected value of Y when the X's are set to 0, β_j is the regression coefficient for each corresponding predictor variable, X_j, and ε is the error of the prediction. Note that $Y - \varepsilon = Y'$ represents the expected value of Y, $E(Y|X_1, X_2, ..., X_n)$. This is also known as the conditional mean.

This multivariate model is useful when the response variable is continuous, but is not appropriate for dichotomous response variables, as is the case when Y is presence(1)/absence(0) of ROP. As it is, the previous model would not produce values restricted to 1 or 0 as we desire. Many uninterpretable values between 0 and 1 and greater than 1 could be obtained. To prevent this from happening, we resort to the model of the logistic distribution.

The logistic regression model indirectly models the response variable based on probabilities associated with the values of Y. We will use $\pi(\mathbf{x})$ to represent the probability that Y=1, which is the presence of ROP.[3] Similarly, we will define 1-$\pi(\mathbf{x})$ to be the probability that Y=0, which is absence of ROP. These probabilities are written in the following form:

$$\pi(\mathbf{x}) = P(Y = 1|X_1, X_2, ..., X_n)$$

$$1 - \pi(\mathbf{x}) = P(Y = 0|X_1, X_2, ..., X_n)$$

In Equation (1) we use the model for the natural logarithm of the odds (log-odds) to favor $Y = 1$.

$$\ln \frac{P(Y = 1|X_1, X_2, ..., X_n)}{1 - P(Y = 1|X_1, X_2, ..., X_n)} = \ln \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \beta_0 + \sum_{j=1}^{n} \beta_j X_j \quad (1)$$

Using the inverse of the logit transformation of Equation (1) we arrive at the following:

$$P(Y = 1|X_1, X_2, ..., X_n) = \frac{e^{\beta_0 + \sum_{j=1}^{n} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{n} \beta_j X_j}} = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{n} \beta_j X_j)}}.$$

[3]Note that the $\mathbf{x}$ in the expression, $\pi(\mathbf{x})$, is a vector representing the set of the independent predictor variables, $X_1, X_2, ..., X_n$.

Thus we have constructed a logistic regression model that bounds the conditional mean between 0 and 1.

Now, we will fit the logistic regression model to the data. First, we must establish a technique for estimating the parameters. The method of parameter estimation is maximum likelihood. We will construct the likelihood function, which expresses the probability of the observed data as a function of the unknown parameters. Then, we will obtain the likelihood estimators of these parameters which maximize the likelihood function. In the process, we will have selected the estimators which predict the observed data most closely.

For a set of observations in the data $(\mathbf{x}_i, y_i)$,[5] the contribution to the likelihood function is $\pi(\mathbf{x}_i)$, where $y_i = 1$, and $1 - \pi(\mathbf{x}_i)$, where $y_i = 0$.[6] The following equation results for the contribution (call it $\zeta(\mathbf{x}_i)$) to the likelihood function for the observation, $(\mathbf{x}_i, y_i)$:

$$\zeta(\mathbf{x}_i) = \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)^{1-y_i}].$$

This equation accounts for only one set of observations. The observations are assumed to be independent of each other so we can multiply their likelihood contributions to obtain the complete likelihood function. The result is given in Equation (2):

$$l(B) = \prod_{i=1}^{n} \zeta(\mathbf{x}_i) \quad (2)$$

where B is the collection of parameters $\beta_0, \beta_1, ..., \beta_j$ and $l(B)$ is the likelihood function of B.

Maximum likelihood estimates (MLE's) can be obtained by calculating the B which maximizes $l(B)$. However, to simplify the mathematics, we will take the logarithm of Equation (2) before finding the value B which maximizes the likelihood function. As shown in Equation (3), $L(B)$ denotes the log likelihood expression.

$$L(B) = \ln [l(B)] = \sum_{i=1}^{n} y_i \ln [\pi(\mathbf{x}_i)] + (1 - y_i) \ln [1 - \pi(\mathbf{x}_i)] \quad (3)$$

We employ the techniques of calculus to determine the value of B that maximizes $L(B)$. This is done by differentiating Equation (3) with respect to $\beta_0, \beta_1, ..., \beta_j$ and setting the resulting derivatives equal to zero. These equations are called likelihood equations, and there are $j + 1$ such equations. They are of the following form:

[5]Note that $\mathbf{x}_i$ is still a vector of the predictor variables, $X_1, X_2, ..., X_n$ for the i^{th} subject.

[6]The expression $\pi(\mathbf{x}_i)$ denotes the value of $\pi(\mathbf{x})$ evaluated at $\mathbf{x}_i$.

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0 \quad [7]$$

for the intercept, β_0, and

$$\sum_{i=1}^{n} x_{ik}[y_i - \pi(\mathbf{x}_i)] = 0 \quad , for \;\; k = 1, 2, ..., j$$

for the predictor variables, $\beta_1, \beta_2, ..., \beta_j$.

The solution to the likelihood equations is the maximum likelihood estimate $\hat{B}$. The solution can be solved for by using computer programs. In particular, SAS will be used to perform the logistic regression analysis of the data for this project and will calculate the maximum likelihood estimates. Therefore, the remainder of this paper is devoted to the analysis of the logistic regression model parameters estimated by SAS.

5 The Model with the Data

The following question posed by Hosmer and Lemeshow outlines the proposed method of testing the significance of predictor variables included in the model: "Does the model that includes the variable in question tell us more about the outcome (or response) variable than does a model that does not include that variable?"[8] To do so, we will use the likelihood ratio test.

The likelihood ratio is given in the following equation:

$$D = -2\ln[l(B)]$$

This statistic, D, is known as the deviance. First, we want to determine the deviance of the model without any of the predictor variables (i.e. with the intercept only), and then compare this value with that of the model consisting of different combinations of variables. The deviance always decreases with the addition of more variables, but the more it decreases, the more that particular predictor variable is related to the response variable. As we add variables, we can evaluate the p-value of the deviance, which tests for the significance of that particular combination of predictor variables. A low p-value[9] justifies the rejection of the null hypothesis, which is that all of the beta coefficients are equal to zero (i.e. the all of the predictor variables are independent of the response variable). The

[7] Note that rearranging this equation yields that the sum of the observed values of y is equal to the sum of the predicted values.

[8] Hosmer, op. cit.

[9] Generally, we accept p-values as significant at the $\alpha = 0.05$ level.

rejection of the null hypothesis means that the variables included in the model are significant.

The deviance for the model containing the intercept only is 316.483. We compare this value to the deviance of the model containing different covariates. The deviance for the model with only the risk factor intraventricular hemorrhage (IVH) is 302.630 with p less than 0.0001. Alone, this predictor variable appears to explain a significant component of presence of ROP. The deviance for the model with only duration of mechanical ventilation is 245.223 with p less than 0.0001. We see that the univariate model for duration of mechanical ventilation predicts ROP better than does the one for IVH. The deviance for the model with only hypertension in the mother is 316.011 with $p = 0.4923$. The difference between the deviance of the intercept only and this risk factor, 0.472, indicates that hypertension in the mother is a poor predictor variable for the presence of ROP. Also, judging by the p-value, we can conclude that the model is better without this covariate.

Other univariate models that yield low deviances (less than 280) are the ones with weight, gestational age, O_2 dependence at 28 days, O_2 dependence at 60 days, postnatal steroids, and number of blood tranfusions. The univariate models that yield high deviances (greater than 314), are the ones containing Apgar score less than 7 at one minute, vaginal birth, breast feeding, gender, and age of mother. Computing deviances for univariate models is not completely reliable. It does not take into account how the predictor variables will interact with each other. It is suspected that some variables, when modeled with others, will become stronger predictors and some weaker. However, it provides a starting place to get general ideas about the predictor variables. We will take this information into consideration when we are trying to determine which of the factors affect the presence of ROP.

We now have an idea which of the risk factors are good predictors and which are not, but we don't yet know how the variables interact with each other. Backwards regression is the method we will use to find the overall best model. That is, we will run a model in SAS containing all of the predictor variables and then analyze the p-values of the predictor variables to determine which of them should be removed. We will also note the deviance of the model, along with its associated p-value, everytime variables are removed in order to further determine the significance of a given combination of independent variables.

The initial deviance of the complete model (with all of the predictor variables) is 145.722 with p less than 0.0001. This is the lowest attainable deviance, since all of the variables are included, but most of the predictor variables don't seem to be affecting the presence of ROP. The p-values of the predictor variables range from 0.0033 to 0.9988. To refine the model, we remove several of the factors exhibiting the highest p-values then reevaluate the model. The resulting deviance never seems to be affected too adversely with the removal of one or two variables (that is, of course, the variables with the highest p-values), and the p-values of the

remaining factors seem to get lower and more significant.

We repeat the removal process until we get to the point where the remaining predictor variables are all significant at the $\alpha = 0.05$ level. Gestational age, duration of mechanical ventilation, multiple gestation, and Therapeutic Intervention Scoring System (TISS score) greater than 20 are the remaining risk factors. Still, some significant factors may have been missed, since they were removed in a certain order. We must check to see if there are any risk factors or combinations of risk factors that can be added back in with siginficant p-values without altering the deviance much. The risk factor, other surgeries, can be added in with a siginficant p-value, lowering the deviance by 4.5. Therefore, we will add this factor to the model. No other risk factor or combination of risk factors can be added to the model and still be significant. However, there are a couple of them that come very close.

The predictor variable, surfactant, has a p-value of 0.0774 and decreases the deviance by 3.4. There has been lots of research done on the effects of surfactant on premature infants. The medical literature confirms that use of surfactant has a positive correlation with the incidence of ROP.[10] Therefore, we will include surfactant in the model. The risk factor breast milk has a p-value of 0.1148 while changing the devaince by 2.7 when added into the model with the other six factors. There have been no studies confirming that breast milk relates to the incidence of ROP, so we do not feel inclined to include it in the model.

The model including the following risk factors: Gestational age, duration of mechanical ventilation, multiple gestation, TISS score greater than 20, other surgeries, and surfactant is the overall best model. The deviance, p-values, and coefficient estimates can be observed from SAS.

6 Conclusion

The overall best logistic regression model for this set of data emphasizes the risk factors that correlate the most with the incidence of ROP. These factors are clincally important because they alert the physicians to the neonates most likely to contract ROP and help them assess the condition of the patient. It is interesting to note that the overall best logistic regression model does not include the risk factor weight. Most of the past research studies of ROP include weight as one of the primary predictors. The logistic regression model shows that weight correlates extremely well with ROP by itself, but not when other variables are included in the model. This is because weight is related to gestational age. If we replace gestational age with weight in the model, then we see that it is almost significant. But, when we place gestational age back in the model, then weight

[10]See, for example, Walther F. J.; et. al. One-year follow-up of 66 premature infants weighing 500-699 grams treated with a single dose of synthetic surfactant or air placebo at birth: results of a double-blind trial. J. Pediatr. 1995; 126(pt 2):13-19.

becomes an insignificant predictor. Gestational age is a much better predictor than weight. This is due to the fact that weight takes into consideration the size of the neonate's parents whereas gestational age takes only the severity of prematurity into account.

It is noteworthy that multiple gestation correlates to the incidence of ROP. No one has ever found this correlation before, and there is no obvious reason why this factor correlates well when other variables are in the model, but is a poor indicator when it is alone. Another interesting finding is the potential correlation of breast milk with the incidence of ROP. This predictor variable was almost included in the model. Breast milk might be a strong predictor, but this research project did not take into account how much breast milk the infant received. It includes infants that received breast milk only once in the same category with those who received it throughout their entire hospital stay. Medical research has shown the potential protective effects of breast milk on premature infants[11], but has never been studied with the incidence of ROP before. Further investigation on the protective effects of breast milk with the incidence of ROP is suggested.

[11]Johnson L.; et. al. Does breast milk-taurine protect against retinopathy of prematurity. Pediatr Res. 1985; 19(pt 2):347A.

7 Appendix

Risk factors for Retinopathy of Prematurity (ROP).

Risk Factor (Number of Categories)	**Description**
Weight (5)	weight of the infant at birth
Gestational Age (5)	number of gestation months before birth
O_2 for 28 days (2)	whether the neonate required oxygen for 28 days or more
Mechanical Ventilation (3)	whether the infant was ever on the mechanical ventilator
O_2 dependent at 60 days (2)	if the infant was dependent on oxygen after the first 60 days of life
Steroids (3)	whether the infant received steroids throughout hospital stay
Surfactant (2)	whether the infant received artificial surfactant throughout hospital stay
IVH (2)	if the infant contracted an intraventricular hemorrhage
Other Surgeries (2)	if the neonate had any surgery other than the standard ones
HCT less than 43 at birth (2)	whether the infants hematocrit level was less than 43 at birth
Number of blood tranfusions (4)	how many blood transfusions the neonate received
Multiple Gestation (2)	whether the infant was single
Apgar score less than 7 at 1 minute (2)	if the infant's Apgar score was less than 7 one minute after birth
Apgar score less than 7 at 5 minutes (2)	if the infant's Apgar score was less than 7 five minutes after birth
Positive blood culture (2)	if an episode of sepsis was confirmed in the infant
Bacterial culture (2)	if the blood infection was bacterial (contingent on positive blood culture)
Fungal culture (2)	if the blood infection was fungal (contingent on positive blood culture)
Vaginal birth (2)	whether the delivery was vaginal or c-section
TPN (3)	whether the neonate received total parenteral nutrition
Hypertension in mother (2)	if the mother experienced hypertension during delivery
NEC (5)	if the infant contracted necrotizing enterocolitis

Risk Factor	**Description**
Race (4)	the race of the infant
Breast milk (2)	whether human breast milk was received during the hospital stay
TISS score (2)	whether the infant attained a TISS (Therapeutic Intervention Scoring System) score greater than 20
Gender (2)	the gender of the infant
Max pO_2 greater than 80 torr. (2)	if the neonate acquired a maximum pO_2 level greater than 80 torr.
O_2 saturation greater than 95% (2)	if the neonate attained an O_2 saturation level greater than 95%
Age of mother (3)	the age of the infant's mother at delivery
Prenatal Care (2)	if the infant's mother received prenatal care