Paper 360-2008

# Convergence Failures in Logistic Regression

Paul D. Allison, University of Pennsylvania, Philadelphia, PA

## ABSTRACT

A frequent problem in estimating logistic regression models is a failure of the likelihood maximization algorithm to converge. In most cases, this failure is a consequence of data patterns known as complete or quasi-complete separation. For these patterns, the maximum likelihood estimates simply do not exist. In this paper, I examine how and why complete or quasi-complete separation occur, and the effects they produce in output from SAS® procedures. I then describe and evaluate several possible solutions.

## INTRODUCTION

Anyone with much practical experience using logistic regression will have occasionally encountered problems with convergence. Such problems are usually both puzzling and exasperating. Most researchers do not have a clue as to why certain models and certain data sets lead to convergence difficulties. And for those who do understand the causes of the problem, it is often unclear whether and how the problem can be fixed.

In this paper, I explain why numerical algorithms for maximum likelihood estimation of the logistic regression model sometimes fail to converge, and I consider a number possible solutions. I also look at how several SAS procedures handle the problem. This paper is a revised and updated version of Allison (2004).

## ML ESTIMATION OF THE LOGISTIC REGRESSION MODEL

I begin with a review of the logistic regression model and maximum likelihood estimation its parameters. For further details, see Allison (1999). For a sample of $n$ cases ($i$=1,…,$n$), we have data on a dummy dependent variable $y_i$ (with values of 1 and 0) and a column vector of explanatory variables $\mathbf{x}_i$ (including a 1 for the intercept term). The logistic regression model states that

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1+\exp(-\boldsymbol{\beta}\mathbf{x}_i)} \tag{1}$$

where $\boldsymbol{\beta}$ is a row vector of coefficients. Equivalently, the model may be written in "logit" form:

$$\ln\left[\frac{\Pr(y_i = 1 \mid \mathbf{x}_i)}{\Pr(y_i = 0 \mid \mathbf{x}_i)}\right] = \boldsymbol{\beta}\mathbf{x}_i \,. \tag{2}$$

Assuming that the $n$ cases are independent, the log-likelihood function for this model is

$$\ell(\boldsymbol{\beta}) = \boldsymbol{\beta}\sum_i \mathbf{x}_i y_i - \sum_i \ln[1+\exp(\boldsymbol{\beta}\mathbf{x}_i)] \tag{3}$$

The goal of maximum likelihood estimation is to find a set of values for $\boldsymbol{\beta}$ that maximize this function. One well-known approach to maximizing a function like this is to differentiate it with respect to $\boldsymbol{\beta}$, set the derivative equal to 0, and then solve the resulting set of equations. The first derivative of the log-likelihood is

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i \hat{y}_i \tag{4}$$

where $\hat{y}_i$, the predicted value of $y$, is given by

$$\hat{y}_i = \frac{1}{1+\exp(-\boldsymbol{\beta}\mathbf{x}_i)} \,. \tag{5}$$

The next step is to set the derivative equal to 0 and solve for $\boldsymbol{\beta}$:

$$\sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i \hat{y}_i = \mathbf{0} \tag{6}$$

Because $\boldsymbol{\beta}$ is a vector, (6) is actually a set of equations, one for each of the parameters to be estimated. These equations are identical to the "normal" equations for least-squares linear regression, except that by (5) $\hat{y}_i$ is a non-linear function of the $\mathbf{x}_i$'s rather than a linear function.

For some models and data (e.g., "saturated" models), the equations in (6) can be explicitly solved for the ML estimator $\hat{\boldsymbol{\beta}}$. For example, suppose there is a single dichotomous *x* variable, so that the data can be arrayed in a $2 \times 2$ table, with observed cell frequencies $f_{11}$, $f_{12}$, $f_{21}$, and $f_{22}$. Then the ML estimator of the coefficient of *x* is given by the logarithm of the "cross-product ratio":

$$\hat{\beta} = \ln\left(\frac{f_{11}f_{22}}{f_{21}f_{22}}\right). \tag{7}$$

For most data and models, however, the equations in (6) have no explicit solution. In such cases, the equations must be solved by numerical methods, of which there are many. The most popular numerical method is the Newton-Raphson algorithm. Let $\mathbf{U}(\boldsymbol{\beta})$ be the vector of first derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$ and let $\mathbf{I}(\boldsymbol{\beta})$ be the matrix of second derivatives. That is,

$$\begin{aligned} \mathbf{U}(\boldsymbol{\beta}) &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i \hat{y}_i \\ \mathbf{I}(\boldsymbol{\beta}) &= \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_i \mathbf{x}_i \mathbf{x}_i' \hat{y}_i(1-\hat{y}_i) \end{aligned} \tag{8}$$

The vector of first derivatives $\mathbf{U}(\boldsymbol{\beta})$ is called the *gradient* while the matrix of second derivatives $\mathbf{I}(\boldsymbol{\beta})$ is called the *Hessian*. The Newton-Raphson algorithm is then

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \mathbf{I}^{-1}(\boldsymbol{\beta}_j)\mathbf{U}(\boldsymbol{\beta}_j) \tag{9}$$

where **I**-1 is the inverse of **I**.

To operationalize this algorithm, a set of starting values $\boldsymbol{\beta}_0$ is required. Choice of starting values is not critical; usually, setting $\boldsymbol{\beta}_0 = \mathbf{0}$ works fine. The starting values are substituted into the right-hand side of (9), which yields the result for the first iteration, $\boldsymbol{\beta}_1$. These values are then substituted back into the right hand side, the first and second derivatives are recomputed, and the result is $\boldsymbol{\beta}_2$. The process is repeated until the maximum change in each parameter estimate from one iteration to the next is less than some criterion, at which point we say that the algorithm has converged. Once we have the results of the final iteration, $\hat{\boldsymbol{\beta}}$, a byproduct of the Newton-Raphson algorithm is an estimate of the covariance matrix of the coefficients, which is just $-\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$. Estimates of the standard errors of the coefficients are obtained by taking the square roots of the main diagonal elements of this matrix.

## WHAT CAN GO WRONG?

A common problem in maximizing a function is the presence of local maxima. Fortunately, such problems cannot occur with logistic regression because the log-likelihood is globally concave, meaning that the function can have at most one maximum (Amemiya 1985). Unfortunately, there are many situations in which the likelihood function has *no* maximum, in which case we say that the maximum likelihood estimate does not exist. Consider the set of data on 10 observations in Table 1.

**Table 1. Data Exhibiting Complete Separation.**

| *x* | *y* |
|---|---|
| -5 | 0 |
| -4 | 0 |
| -3 | 0 |
| -2 | 0 |
| -1 | 0 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |

For these data, it can be shown that the ML estimate of the intercept is 0. Figure 1 shows a graph of the log-likelihood as a function of the slope "beta".

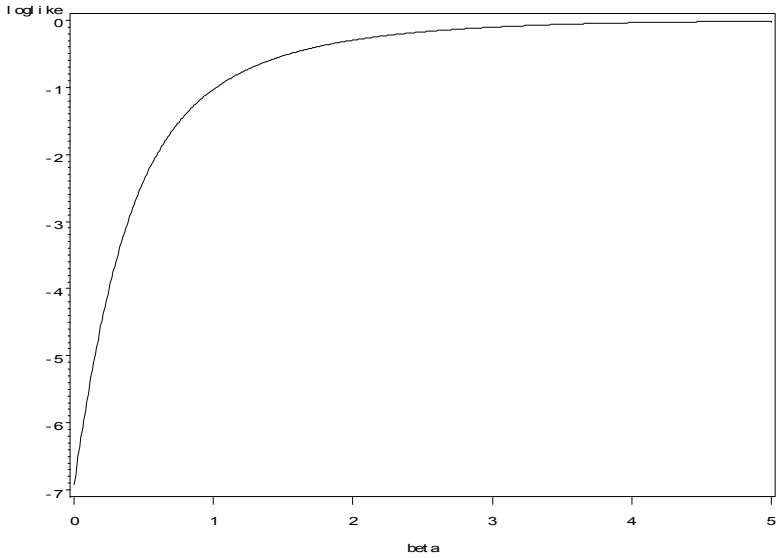


**Figure 1. Log-likelihood as a function of the slope under complete separation**

It is apparent that, although the log-likelihood is bounded above by 0, it does not reach a maximum as beta increases. We can make the log-likelihood as close to 0 as we choose by making beta sufficiently large. Hence, there is no maximum likelihood estimate.

This is an example of a problem known as *complete separation* (Albert and Anderson 1984), which occurs whenever there exists some vector of coefficients **b** such that $y_i = 1$ whenever $\mathbf{bx}_i > 0$ and $y_i = 0$ whenever $\mathbf{bx}_i \leq 0$. In other words, complete separation occurs whenever a linear function of **x** can generate perfect predictions of *y*. For our hypothetical data set, a simple linear function that satisfies this property is 0 + 1(*x*). That is, when *x* is greater than 0, *y*=1, and when *x* is less than or equal to 0, *y*=0.

A related problem is known as *quasi-complete separation*. This occurs when (a) there exists some coefficient vector **b** such that $\mathbf{bx}_i \geq 0$ whenever $y_i = 1$, and $\mathbf{bx}_i \leq 0$ whenever $y_i = 0$, and equality holds for at least one case in each category of the dependent variable. Table 2 displays a data set that satisfies this condition.

**Table 2. Data Exhibiting Quasi-Complete Separation.**

| *x* | *y* |
|---|---|
| -5 | 0 |
| -4 | 0 |
| -3 | 0 |
| -2 | 0 |
| -1 | 0 |
| 0 | 0 |
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |

What distinguishes this data set from the previous one is that there are two additional observations, each with *x* values of 0 but having different values of *y*.

The log-likelihood function for these data, shown in Figure 2, is similar in shape to that in Figure 1. However, the asymptote for the curve is not 0, but a number that is approximately -1.39. In general, the log-likelihood function for quasi-complete separation will not approach 0, but some number lower than that. In any case, the curve has no maximum so, again, the maximum likelihood estimate does not exist.

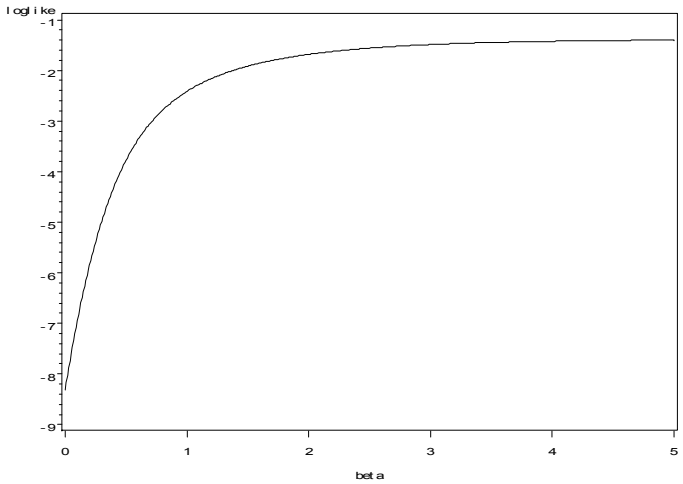


**Figure 1. Log-likelihood as a function of the slope, quasi-complete separation.**

Of the two conditions, complete and quasi-complete separation, the latter is far more common. It most often occurs when an explanatory variable *x* is a dummy variable and, for one value of *x*, either every case has the event *y*=1 or every case has the event *y*=0. Consider the following 2 × 2 table:

| | | *y* = 1 | *y* = 0 |
|---|---|---|---|
| *x* | 1 | 5 | 0 |
| | 0 | 15 | 10 |

If we form the linear function $c = 0 + (1)\,x$, we have $c \geq 0$ when $y=1$ and $c \leq 0$ when $y=0$. Further, for all the cases in the second row, $c = 0$ for both values of $y$. So the conditions of quasi-complete separation are satisfied.

To get some intuitive sense of why this leads to non-existence of the maximum likelihood estimator, consider equation (7) which gives the maximum likelihood estimator of the slope coefficient for a 2 × 2 table. For our quasi-complete table, that would be

$$\hat{\beta} = \ln\left(\frac{5\times 10}{15\times 0}\right).$$

But this is undefined because there is a zero in the denominator. The same problem would occur if there were a zero in the numerator because the logarithm of zero is also undefined. If the table is altered to read

| | | *y* | |
|---|---|---|---|
| | | 1 | 0 |
| 1 | | 5 | 0 |
| 0 | | 0 | 10 |

then there is *complete* separation with zeros in both the numerator and the denominator.

So the general principle is evident: Whenever there is a zero in any cell of a $2 \times 2$ table, the maximum likelihood estimate of the logistic slope coefficient does not exist. This principle also extends to multiple independent variables:

*For any dichotomous independent variable in a logistic regression, if there is a zero in the $2 \times 2$ table formed by that variable and the dependent variable, the ML estimate for the regression coefficient does not exist*.

This is by far the most common cause of convergence failure in logistic regression. Obviously, it is more likely to occur when the sample size is small. Even in large samples, it will frequently occur when there are extreme splits on the frequency distribution of either the dependent or independent variables. Consider, for example, a logistic regression predicting whether a person has some disease whose overall prevalence is less than 1 in 1000 . Suppose further, that the explanatory variables include a set of seven dummy variables representing different age categories. Even if the sample contained 20,000 cases, we could reasonably expect that for at least one of the categories, no one would have the disease.

## BEHAVIOR OF THE NEWTON-RAPHSON ALGORITHM UNDER SEPARATION

We just saw that when there are explicit formulas for the maximum likelihood estimate and there is either complete or quasi-complete separation, the occurrence of zeros in the formulas prevents computation. What happens when the Newton-Raphson algorithm is applied to data exhibiting either kind of separation? That depends on the particular implementation of the algorithm. The classic behavior is this: at each iteration, the parameter estimate for the variable (or variables) with separation gets larger in magnitude. Iterations continue until the fixed iteration limit is exceeded. At whatever limit is reached, the parameter estimate is large and the estimated standard error is extremely large. If separation is complete, the log-likelihood will be reported as zero.

What actually happens depends greatly on how the algorithm is implemented. To determine how several SAS procedures handle complete and quasi-complete separation, I tried estimating logistic regression models for the data sets in Tables 1 and 2. Results are summarized in Table 3. What follows is a detailed discussion of each of the headings in this table.

### WARNING MESSAGES

Ideally, the program should detect the separation and issue a clear warning message to the user. In their classic paper on separation in logistic regression, Albert and Anderson (1984) proposed an "empirical" method that has been implemented in PROC LOGISTIC. It has the following steps:

(a) If the convergence criterion is satisfied within eight iterations, conclude that there is no problem.

(b) For all iterations after the eighth, compute the predicted probability of the observed response for each observation, which is given by

$$\hat{y}_i = \frac{1}{1+\exp[(2y_i - 1)\hat{\boldsymbol{\beta}}\mathbf{x}_i]}$$

If the predicted probability is one for all observations, conclude that there is complete separation and stop the iterations.

(c) If the probability of the observed response is large ($\geq$.95) for some observations (but not all), examine the estimated standard errors for that iteration. If they exceed some criterion, conclude that there is quasi-complete separation and stop the iteration.

The check for complete separation is very reliable, but the check for quasi-complete separation is less so. For more

reliable checks of quasi-complete separation, methods based on linear programming algorithms have been proposed by Albert and Anderson (1984) and Santner and Duffy (1986

When the data set in Table 1 was used with PROC LOGISTIC, it printed the message:

```
                        Complete separation of data points detected.
WARNING: The maximum likelihood estimate does not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
         Results shown are based on the last maximum likelihood iteration.
         Validity of the model fit is questionable.
```

For quasi-complete separation, the message was

```
                      Quasicomplete separation of data points detected.
WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
         Results shown are based on the last maximum likelihood iteration.
         Validity of the model fit is questionable.
```

While PROC LOGISTIC (and its close cousin, SURVEYLOGISTIC) came close to the ideal, other procedures left much to be desired with regard to detection and warnings. GENMOD and PROBIT gave no warning message for the complete separation example. For the quasi-complete example, both GENMOD and PROBIT gave the somewhat cryptic message "Negative of Hessian not positive definite." For both complete- and quasi-complete separation, CATMOD marked non-convergent coefficient estimates with the symbol # and said that they were "regarded to be infinite." QLIM gave no warning messages in either case.

#### FALSE CONVERGENCE

Strictly speaking, the Newton-Raphson algorithm should not converge under either complete or quasi-complete separation. Nevertheless, the only procedure that exhibited this classic behavior was PROBIT in the case of complete separation. No matter how high I raised the maximum number of iterations, PROBIT failed to achieve convergence. QLIM and CATMOD, on the other hand, claimed to have reached convergence for both examples. LOGISTIC and SURVEYLOGISTIC stopped the iterations once separation had been detected. GENMOD reported convergence for the complete separation example, but not for the quasi-complete example.

Why do some procedures claim convergence for estimates that do not exist? The reason is that their iteration algorithms have been modified to "improve" convergence. GENMOD, for example, uses a Newton-Raphson algorithm that performs a line search at each step in order to stabilize its convergence properties. In most cases, this improves the rate of convergence. However, if the likelihood cannot be improved along the current Newton step, the algorithm returns the current parameter values as the updated values and therefore determines that the convergence criterion has been met (Johnston 2002). Unfortunately, the combination of apparent convergence and lack of clear warning messages in many procedures means that some users are likely to be misled about the validity of their parameter estimates.

**Table 3.** Performance of Logistic Regression PROCs under Complete and Quasi-Complete Separation.

| | **Warning Messages** | | **False Convergence** | | **Report Estimates** | | **LR Statistics** | |
|---|---|---|---|---|---|---|---|---|
| | Comp | Quasi | Comp | Quasi | Comp | Quasi | Comp | Quasi |
| LOGISTIC | C | C | | | * | * | | |
| SURVEYLOGISTIC | C | C | | | * | * | | |
| GENMOD | | A | * | | * | * | * | * |
| CATMOD | A | A | * | * | * | * | | |
| PROBIT | | A | | | * | * | | |
| QLIM | | | * | * | * | * | | |

Note: C=clear warning, A=ambiguous warning

#### REPORTING OF PARAMETER ESTIMATES AND STANDARD ERRORS

Some software packages (e.g., SPSS and STATA) do a reasonably good job of detecting and warning about separation, but then fail to report any parameter estimates or standard errors. This might seem sensible since non-convergent estimates are essentially worthless as parameter estimates. However, they may still serve a useful diagnostic purpose in determining which variables have complete or quasi-complete separation. All the SAS procedures examined here do report parameter estimates and standard errors for the final iteration.

#### LIKELIHOOD RATIO STATISTICS

PROC GENMOD can report optional likelihood-ratio chi-square tests for each of the coefficients in the model. Unlike Wald chi-squares, which are essentially useless under complete or quasi-complete separation, the likelihood ratio test is still a valid test of the null hypothesis that a coefficient is equal to 0. Thus, even if a certain parameter cannot be estimated, it may still be possible to judge whether it is significantly different from 0.

## DIAGNOSIS OF SEPARATION PROBLEMS

We are now in a position to make some recommendations about how the statistical analyst should approach the detection of problems of complete or quasi-complete separation. If you are using software that gives clear diagnostic messages (e.g., PROC LOGISTIC), then half the battle is won. But there is still a need to determine which variables are causing the problem, and to get a better sense of the nature of the problem.

The second step (or the first step with programs that do not give good warning messages) is to carefully examine the estimated coefficients and their standard errors. Variables with non-existent coefficients will invariably have large parameter estimates, typically greater than 5.0, and huge standard errors, producing Wald chi-square statistics that are near 0. If any of these variables is a dummy (indicator) variable, the next step is to construct the $2 \times 2$ table for each dummy variable with the dependent variable. A frequency of zero in any single cell of the table implies quasi-complete separation. Less commonly, if there are two diagonally opposed zeros in the table, the condition is complete separation.

Once you have determined which variables are causing separation problems, it is time to consider possible solutions. The potential solutions are somewhat different for complete and quasi-complete separation, so I will treat them separately. I begin with the more common problem of quasi-complete separation.

## SOLUTIONS FOR QUASI-COMPLETE SEPARATION

#### DELETION OF PROBLEM VARIABLES

In practice, the most widely used method for dealing with quasi-complete separation is simply to delete from the model any variables whose coefficients did not converge. *I do not recommend this method*. If a variable has quasi-complete separation with the dependent variable, it is reasonable to suppose that that variable has a strong (albeit, non-infinite) effect on the dependent variable. Deleting variables with strong effects will certainly obscure the effects of those variables, and is also likely to bias the coefficients for other variables in the model.

#### COMBINING CATEGORIES

As noted earlier, the most common cause of quasi-complete separation is a dummy predictor variable such that, for one level of the variable, either every observation has the event or no observation has the event. For those cases in which the problem variable is one of a set of variables representing a single categorical variable, the problem can often be easily solved by combining categories. For example, suppose that marital status has five categories: never married, currently married, divorced, separated, and widowed. This variable could be represented by four dummy variables, with currently married as the reference category. Suppose, further, that the sample contains 50 persons who are divorced but only 10 who are separated. If the dependent variable is 1 for employed and 0 for unemployed, it is quite possible that all 10 of the separated persons would be employed, leading to quasi-complete separation. A natural and simple solution is to combine the divorced and separated categories, turning two dummy variables into a single dummy variable.

Similar problems often arise when a quantitative variable, like age, is divided into a set of categories, with dummy variables for all but one of the categories. Although this can be a useful device for representing non-linear effects, it can easily lead to quasi-complete separation if the number of categories is large and the number of cases within some categories is small. The solution is to use a smaller number of categories, or perhaps revert to the original

quantitative representation of the variable.

If the dummy variable represents an irreducible dichotomy, like sex, then this solution is clearly not feasible. However, there is another simple method that often provides a very satisfactory solution.

**DO NOTHING AND REPORT LIKELIHOOD RATIO CHI-SQUARES**

Just because maximum likelihood estimates do not exist for some coefficients because of quasi-complete separation, they may still exist for other variables in the logistic regression model. In fact, if one leaves the offending variables in the model, the coefficients, standard errors, and test statistics for the remaining variables are still valid maximum likelihood estimates. Thus, one attractive strategy is just to leave the problem variables in the model. The coefficients for those variables could be reported as $+\infty$ or $-\infty$. The standard errors and Wald statistics for the problem variables will certainly be incorrect but, as noted above, likelihood ratio tests for the null hypothesis that the coefficient is zero are still valid. If likelihood-ratio statistics are not available as options in the procedure, they can be obtained by fitting the model with and without each problem variable, then taking twice the positive difference in the log-likelihoods.

If the problem variable is a dummy variable, then the estimates obtained for the *other* variables have a special interpretation. They are the ML estimates for the subsample of cases that fall into the category of the dummy variable in which observations differ on the dependent variable. For example, suppose that the dependent variable is whether or not a person smokes cigars. A dummy variable for sex is included in the model, but none of the women smoke cigars, producing quasi-complete separation. If sex is left in the model, the coefficients for the remaining variables (e.g., age, income, education) represent the effects of those variables among men only. (This can easily be verified by actually running the model for men only). The advantage of doing it in the full sample with sex as a covariate is that one also gets a test of the sex effect (using the likelihood ratio chi-square) while controlling for the other predictor variables.

**EXACT INFERENCE**

As previously mentioned, problems of separation are most likely to occur in small samples and/or when there is an extreme split on the dependent variable. Of course, even without separation problems, maximum likelihood estimates may not have good properties in small samples. One possible solution is to abandon maximum likelihood entirely and do exact logistic regression. This method, originally proposed by Cox (1970), was not computationally feasible until relatively recently, and is now available in PROC LOGISTIC.

Exact logistic regression is designed to produce exact *p*-values for the null hypothesis that a specified predictor variable has a coefficient of 0, conditional on all the other predictors. These *p*-values, based on permutations of the data rather than on large-sample chi-square approximations, are essentially unaffected by complete or quasi-complete separation. The coefficient estimates reported with this method are usually conditional maximum likelihood estimates, but even these may not exist when there is separation. In that event, PROC LOGISTIC reports *median unbiased estimates* for the problem coefficients. If the true value is $\beta$, a median unbiased estimator $\beta_u$ has the property

$$\Pr(\beta_u \le \beta) \ge 1/2, \qquad \Pr(\beta_u \ge \beta) \ge 1/2$$

Hirji et al. (1989) demonstrated that the median unbiased estimator is generally more accurate than the maximum likelihood estimator for small sample sizes.

I used PROC LOGISTIC to do exact estimation for the data in Tables 1 and 2. For the completely separated data in Table 1, the *p*-value for the coefficient of *x* was .0079. The median unbiased estimate was .7007. For the quasi-completely separated data in Table 2, the *p*-value was .0043 with a median unbiased estimate of .9878. Despite the attractiveness of exact logistic regression, it is essential to emphasize that it is computationally feasible only for rather small samples, or when the number of cases on one the two outcomes is small.

**BAYESIAN ESTIMATION**

In situations where none of the preceding solutions is appropriate, another possibility is to do Bayesian estimation with a prior distribution on the regression coefficients (Hsu and Leonard 1997, Kahn and Raftery 1996). This should soon be possible in release 9.2 of SAS with the BAYES statement in PROC GENMOD. Based on a few, limited tests of this new capability, I have concluded that a non-informative prior does not work satisfactorily in dealing with

separation. On the other hand, when using informative priors, results may depend rather heavily on the mean and variance of the prior distribution.

#### PENALIZED MAXIMUM LIKELIHOOD ESTIMATION

A very promising new method for dealing with separation is penalized maximum likelihood estimation, proposed by Firth (1993) to reduce bias in logistic regression in small samples. Heinze and Schemper (2002) have shown that this method always yields finite estimates of parameters under complete or quasi-complete separation. Their simulation results indicate that these estimates have relatively little bias, even under extreme conditions. In fact, the bias is appreciably less than that found for median unbiased estimates associated with exact logistic regression. And unlike exact logistic regression, penalized maximum likelihood is computationally feasible even for large samples.

Firth's procedure replaces the gradient vector $\mathbf{U}(\boldsymbol{\beta})$ in equations (8) with

$$\mathbf{U}(\boldsymbol{\beta}^*) = \sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i \hat{y}_i - \sum_i h_i \mathbf{x}_i (.5 - \hat{y}_i)$$

where $h_i$ is the $i$'th diagonal element of the "hat" matrix $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$ and $\mathbf{W} = \text{diag}\{\hat{y}_i(1-\hat{y}_i)\}$. Once this replacement is made, the Newton-Raphson algorithm of equation (9) can proceed in the usual way. Standard errors are also calculated as usual by taking the square roots of the diagonal elements of $-\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$. However, Heinze and Schemper (2002) point out that Wald tests based on the standard errors for variables causing separation can be highly inaccurate (as with conventional maximum likelihood). Instead, they recommend chi-square tests based on differences between penalized log-likelihoods.

The penalized maximum likelihood method has been implemented in a macro for SAS (Heinze 1999). Using that macro, I estimated models for the data in Tables 1 and 2. For the data with complete separation (Table 1), the parameter estimate was .936 with a *p*-value of .0032 (based on a penalized likelihood ratio chi-square). For the data with quasi-complete separation (Table 2), the parameter estimate was .853 with a *p*-value of .0037. These estimates and *p*-values are similar to those reported earlier using the exact method with median unbiased estimation. The penalized maximum likelihood method is expected to be incorporated into release 9.2 of PROC LOGISTIC.

## SOLUTIONS FOR COMPLETE SEPARATION

It is fortunate that complete separation is less common than quasi-complete separation because, when it occurs, it is considerably more difficult to deal with. For example, leaving the problem variable in the model is not an option because that makes it impossible to get maximum likelihood estimates for any other variables. And combining categories for dummy variables will not solve the problem either. Exact logistic regression may be useful with small samples and a single predictor variable causing complete separation. But it is not computationally feasible for larger samples, and cannot produce coefficient estimates for any additional predictor variables. That's because the permutation distribution is degenerate when one conditions on a variable causing complete separation. Bayesian estimation may be a feasible solution, but it requires an informative prior distribution on the problem parameters, and results may be sensitive to the choice of that distribution.

With conventional logistic regression, about the only practical approach to dealing with complete separation is to delete the problem variable from the model. That allows one to get estimates for the remaining variables but, as noted earlier, the exclusion of the problem variable could lead to biased estimates for the other variables. If one chooses to go this route, it is also essential to compute and report likelihood ratio chi-squares or exact *p*-values (where feasible) for the bivariate relationship between the problem variable and the dependent variable so that the statistical significance of this relationship can be assessed.

A better solution is to use the penalized likelihood method. As noted above, this will produce approximately unbiased estimates of coefficients, and it will work with more than one predictor.

## EXTENSIONS AND CONCLUSION

In this paper, I have focused entirely on problems of non-convergence with binary logistic regression. But it is important to stress that complete and quasi-complete separation also lead to non-existence of maximum likelihood estimates under other "link" functions for binary dependent variables, including the probit model and the complementary log-log model. For the most part, software treatment of data with separation is the same with these

link functions as with the logit link. The possible solutions I described for the logistic model should also work for these alternative link functions, with one exception: the computation of exact *p*-values is only available for the logit link function.

Data separation can also occur for the unordered multinomial logit model; in fact, complete and quasi-complete separation were first defined in this more general setting (Albert and Anderson 1984). Separation problems can also occur for the cumulative logit model although, to my knowledge, separation has not been rigorously defined for this model. Table 3 displays data with complete separation for a three-valued, ordered dependent variable.

**Table 3. Ordered Data Exhibiting Quasi-Complete Separation.**

| *x* | *y* |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 2 |
| 6 | 2 |
| 7 | 2 |
| 8 | 2 |
| 9 | 3 |
| 10 | 3 |
| 11 | 3 |
| 12 | 3 |

These data could be modified to produce quasi-complete separation by adding a new observation with $x$=5 and $y$=1. PROC LOGISTIC in SAS correctly identifies both of these conditions and issues the same warning messages we saw earlier.

## REFERENCES

Albert, A. and J. A. Anderson (1984) "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71: 1-10.

Albert, James H. and Siddhartha Chib (1993) "Bayesian Analysis of Binary and Polychotomous Data." *Journal of the American Statistical Association* 88: 669-679.

Allison, Paul D. (1999) *Logistic Regression Using SAS: Theory and Practice*. Cary, NC: The SAS Institute.

Allison, Paul D. (2004) "Convergence Problems in Logistic Regression." Pp. 238-252 in Micah Altman, Jeff Gill and Michael P. McDonald (eds.), *Numerical Issues in Statistical Computing for the Social Scientist.* Hoboken, NJ: John Wiley & Sons.

Amemiya, Takeshi (1985) *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Cox, D. R. (1970) *Analysis of Binary Data*. London: Chapman and Hall.

Firth, D. (1993) "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80: 27-38.

Heinze, Georg (1999) "The Application of Firth's Procedure to Cox and Logistic Regression." Technical Report 10/1999. Department of Medical Computer Sciences, Section of Clinical Biometrics, Vienna University. Available at http://www.akh-wien.ac.at/imc/biometrie/programme/fl_en.

Heinze, George and Michael Schemper (2002) "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21: 2409-2419.

Hirji, Karim F., Anastasios A. Tsiatis and Cyrus R. Mehta (1989) "Median Unbiased Estimation for Binary Data." *The American Statistician* 43: 7-11.

Hsu, J.S.J. and T. Leonard (1997) “Hierarchical Bayesian Semiparametric Procedures for Logistic Regression.” *Biometrika* 84: 85-93.

Johnston, Gordon (2002). Personal communication.

Kahn M .J. and A. E. Raftery (1996) “Discharge Rates of Medicare Stroke Patients to Skilled Nursing Facilities: Bayesian Logistic Regression with Unobserved Heterogeneity.” *Journal Of The American Statistical Association* 91: 29-41

Santner, Thomas J. and Diane E. Duffy (1986) “A Note on A. Albert and J. A. Anderson’s Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models.” *Biometrika* 73: 755-758.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Paul D. Allison
Sociology Department, University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6299
Work Phone: 215-898-6717
Fax: 215-573-2081
E-mail: allison@soc.upenn.edu
Web: www.ssc.upenn.edu/~allison