Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2007 *Paper* 26

The Existence of Maximum Likelihood Estimates for the Binary Response Logistic Regression Model

William F. McCarthy[*]

[*]Maryland Medical Research Institute, dr.w.f.mccarthy@gmail.com

The Existence of Maximum Likelihood Estimates for the Binary Response Logistic Regression Model

William F. McCarthy

Abstract

The existence of maximum likelihood estimates for the binary response logistic regression model depends on the configuration of the data points in your data set. There are three mutually exclusive and exhaustive categories for the configuration of data points in a data set: Complete Separation, Quasi-Complete Separation, and Overlap. For this paper, a binary response logistic regression model is considered. A 2 x 2 tabular presentation of the data set to be modeled is provided for each of the three categories mentioned above. In addition, the paper will present an example of a data set whose data points have a linear dependency. Both unconditional maximum likelihood estimation (asymptotic inference) and exact conditional estimation (exact inference) will be considered and contrasted in terms of results. The statistical software package SAS will be used for the binary response logistic regression modeling.

Introduction

The Existence of Maximum Likelihood Estimates for the Logistic Regression Model depends on the configuration of the data points in your data set (Albert and Anderson, 1984; Santner and Duffy, 1985; So, 1995). There are three mutually exclusive and exhaustive categories for the configuration of data points in a data set:

- Complete Separation
- Quasi-Complete Separation
- Overlap

Refer to So (1995) for a nice graphical illustration of these three categories. For this paper, a binary response logistic regression model is considered. A 2 x 2 tabular presentation of the data set to be modeled is provided for each of the three categories mentioned above. In addition, the paper will present an example of a data set whose data points have a linear dependency.

Unconditional maximum likelihood estimation (asymptotic inference) is used when matched data are not considered, provided that the total number of variables in the model is not too large relative to the number of observations (Kleinbaum, 1994). This method of inference is based on maximizing the likelihood function for parameter estimation using the unconditional formula (Kleinbaum, 1994). This is the usual large-sample asymptotic method used by most of the current statistical software packages (Kleinbaum, 1994; Mehta and Patel, 1995).

The existence and uniqueness of maximum likelihood parameter estimates for the logistic regression model depends on the pattern of the data points in the observation space (Albert and Anderson, 1984; Santer and Duffy, 1986; So, 1993).

Complete Separation of data points gives non-unique infinite parameter estimates. Thus, maximum likelihood parameter estimates do not exist. Quasi-Complete Separation of data points also gives non-unique infinite parameter estimates. Thus, maximum likelihood parameter estimates do not exist. Maximum likelihood parameter estimates exist and are unique when there is an Overlap of data points. Complete separation and quasi-complete separation of data points usually occur with small data sets. Complete separation can occur for any type of data, but quasi-complete separation is not likely for quantitative data.

To contrast unconditional maximum likelihood estimation, exact conditional estimation will be considered as well. The theory of exact conditional logistic regression analysis (exact inference) was first proposed by Cox (1970). The computational methods employed in the statistical software package SAS (PROC LOGISTIC) are described in Hirji, Mehta, and Patel (1987), Hirji (1992), and Mehta, Patel, and Senchaudhuri (1992).

Exact conditional inference is based on generating the conditional distribution for the sufficient statistics of the parameters of interest. This distribution is called the permutation or exact conditional distribution. If the sufficient statistic of the β being estimated lies at one extreme of its range, a median unbiased estimate is reported (Hirji, Tsiatis, and Mehta 1989).

Methods

This paper will use both asymptotic and exact inference when modeling the data and will present the SAS output obtained when each of the four data sets are used for modeling. The emphasis of this paper is to show how SAS handles these four data sets when both asymptotic and exact inference is used with respect to a binary response logistic regression modeling. The outcome variable is binary (Mutation; YES, NO) and the covariate is categorical (Drug; EXPOSURE, NON-EXPOSURE). An intercept (Constant) term will be included in the model as well.

Data Sets that will be considered in this paper:

Complete Separation

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	0	14	14
Drug=EXPOSURE	37	0	37
Total	37	14	51

Quasi-Complete Separation

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	12	0	12
Drug=EXPOSURE	25	14	39
Total	37	14	51

Overlap

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	9	3	12
Drug=EXPOSURE	25	14	39
Total	34	17	51

Linear Dependency

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	0	0	0
Drug=EXPOSURE	37	14	51
Total	37	14	51

SAS Output

The SAS output for asymptotic and exact inference when considering the Complete Separation data set is presented below.

Complete Separation

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	**0**	**14**	**14**
Drug=EXPOSURE	**37**	**0**	**37**
Total	**37**	**14**	**51**

```
                                               The LOGISTIC Procedure

                                                 Model Information

                                   Data Set                      WORK.TEST
                                   Response Variable             mutation
                                   Number of Response Levels     2
                                   Number of Observations        51
                                   Model                         binary logit
                                   Optimization Technique        Fisher's scoring


                                                  Response Profile

                                         Ordered                      Total
                                           Value     mutation     Frequency

                                               1            1            14
                                               2            0            37

                                         Probability modeled is mutation=1.


                                              Model Convergence Status

                                    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.


                                                Model Fit Statistics

                                                                    Intercept
                                                     Intercept            and
                                       Criterion          Only     Covariates

                                       AIC              61.945          4.007
                                       SC               63.877          7.871
                                       -2 Log L         59.945          0.007
```

```
                                      Testing Global Null Hypothesis: BETA=0

                              Test                 Chi-Square       DF     Pr > ChiSq

                              Likelihood Ratio        59.9375        1         <.0001
                              Score                   51.0000        1         <.0001
                              Wald                     0.2295        1         0.6319


                         Analysis of Maximum Likelihood Estimates

                                                  Standard        Wald
                  Parameter     DF    Estimate       Error      Chi-Square    Pr > ChiSq
                  Intercept      1      9.7773     35.4872        0.0759        0.7829
                  drug           1    -19.2725     40.2338        0.2295        0.6319


                                               The LOGISTIC Procedure
WARNING: The validity of the model fit is questionable.

                                                Odds Ratio Estimates

                                                    Point          95% Wald
                                   Effect        Estimate      Confidence Limits

                                   drug           <0.001      <0.001    >999.999


                           Association of Predicted Probabilities and Observed Responses

                                 Percent Concordant    100.0    Somers' D    1.000
                                 Percent Discordant      0.0    Gamma        1.000
                                 Percent Tied            0.0    Tau-a        0.406
                                 Pairs                   518    c            1.000


                                      Wald Confidence Interval for Parameters

                                 Parameter      Estimate     95% Confidence Limits

                                 Intercept        9.7773     -59.7764      79.3309
                                 drug           -19.2725     -98.1293      59.5842
                              The LOGISTIC Procedure


                            Exact Conditional Analysis

                             Conditional Exact Tests

                                                       --- p-Value ---
               Effect       Test          Statistic     Exact      Mid

               Intercept    Score           14.0000    0.0001    <.0001
                            Probability    0.000061    0.0001    <.0001
               drug         Score           50.0000    <.0001    <.0001
                            Probability    7.74E-13    <.0001    <.0001


                            Exact Parameter Estimates

                                          95% Confidence
             Parameter     Estimate           Limits          p-Value

             Intercept       2.9807*      1.1991    Infinity     0.0001
             drug           -6.4343*   -Infinity     -4.1539     <.0001

                  NOTE: * indicates a median unbiased estimate.
```

The SAS output for asymptotic and exact inference when considering the Quasi-Complete Separation data set is presented below.

Quasi-Complete Separation

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	**12**	**0**	**12**
Drug=EXPOSURE	**25**	**14**	**39**
Total	**37**	**14**	**51**

```
                              The LOGISTIC Procedure

                                Model Information

                    Data Set                      WORK.TEST
                    Response Variable             mutation
                    Number of Response Levels     2
                    Number of Observations        51
                    Model                         binary logit
                    Optimization Technique        Fisher's scoring
```

```
                                                  Response Profile

                                         Ordered                      Total
                                           Value     mutation     Frequency

                                               1            1            14
                                               2            0            37

                                         Probability modeled is mutation=1.


                                              Model Convergence Status

                                 Quasi-complete 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.

                                                Model Fit Statistics

                                                                    Intercept
                                                     Intercept            and
                                       Criterion          Only     Covariates

                                       AIC              61.945         54.920
                                       SC               63.877         58.784
                                       -2 Log L         59.945         50.920


                                      Testing Global Null Hypothesis: BETA=0

                              Test                 Chi-Square       DF     Pr > ChiSq

                              Likelihood Ratio         9.0242        1         0.0027
                              Score                    5.9376        1         0.0148
                              Wald                     0.0028        1         0.9581


                                     Analysis of Maximum Likelihood Estimates

                                                    Standard          Wald
                    Parameter    DF    Estimate        Error    Chi-Square    Pr > ChiSq

                    Intercept     1    -13.4954        246.0        0.0030        0.9562
                    drug          1     12.9155        246.0        0.0028        0.9581
```

```
                                               The LOGISTIC Procedure
WARNING: The validity of the model fit is questionable.

                                                Odds Ratio Estimates

                                                  Point          95% Wald
                                     Effect    Estimate      Confidence Limits

                                     drug      >999.999      <0.001    >999.999


                           Association of Predicted Probabilities and Observed Responses

                                 Percent Concordant     32.4    Somers' D    0.324
                                 Percent Discordant      0.0    Gamma        1.000
                                 Percent Tied           67.6    Tau-a        0.132
                                 Pairs                   518    c            0.662


                                      Wald Confidence Interval for Parameters

                                  Parameter     Estimate     95% Confidence Limits

                                  Intercept     -13.4954       -495.6        468.6
                                  drug           12.9155       -469.2        495.0


                                               The LOGISTIC Procedure

                                             Exact Conditional Analysis

                                              Conditional Exact Tests

                                                                     --- p-Value ---
                               Effect      Test          Statistic    Exact      Mid

                               Intercept   Score           12.0000   0.0005   0.0004
                                           Probability    0.000244   0.0005   0.0004
                               drug        Score            5.8212   0.0224   0.0166
                                           Probability      0.0117   0.0224   0.0166


                                             Exact Parameter Estimates

                                                          95% Confidence
                             Parameter    Estimate            Limits           p-Value

                             Intercept     -2.8224*   -Infinity     -1.0219     0.0005
                             drug           2.1708*      0.2484    Infinity     0.0233

                                   NOTE: * indicates a median unbiased estimate.
```

The SAS output for asymptotic and exact inference when considering the Overlap data set is presented below.

Overlap

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	**9**	**3**	**12**
Drug=EXPOSURE	**25**	**14**	**39**
Total	**34**	**17**	**51**

```
            The LOGISTIC Procedure

              Model Information

Data Set                     WORK.TEST
Response Variable            mutation
Number of Response Levels    2
Number of Observations       51
Model                        binary logit
Optimization Technique       Fisher's scoring


               Response Profile

      Ordered                      Total
        Value     mutation     Frequency

            1            1            17
            2            0            34

      Probability modeled is mutation=1.


           Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.


             Model Fit Statistics

                                 Intercept
                  Intercept            and
    Criterion          Only     Covariates

    AIC              66.924         68.416
    SC               68.856         72.280
    -2 Log L         64.924         64.416
```

```
                 Testing Global Null Hypothesis: BETA=0

         Test                 Chi-Square       DF     Pr > ChiSq

         Likelihood Ratio         0.5080        1         0.4760
         Score                    0.4904        1         0.4838
         Wald                     0.4838        1         0.4867


                Analysis of Maximum Likelihood Estimates

                                Standard      Wald
Parameter    DF    Estimate        Error    Chi-Square    Pr > ChiSq

Intercept     1     -1.0984       0.6666        2.7148        0.0994
drug          1      0.5186       0.7455        0.4838        0.4867


                           Odds Ratio Estimates

                             Point          95% Wald
                Effect    Estimate      Confidence Limits

                drug         1.680       0.390       7.241


                          The LOGISTIC Procedure

      Association of Predicted Probabilities and Observed Responses

            Percent Concordant     21.8    Somers' D    0.088
            Percent Discordant     13.0    Gamma        0.254
            Percent Tied           65.2    Tau-a        0.040
            Pairs                   578    c            0.544


                 Wald Confidence Interval for Parameters

             Parameter     Estimate     95% Confidence Limits

             Intercept      -1.0984      -2.4050       0.2082
             drug            0.5186      -0.9427       1.9798


                          The LOGISTIC Procedure

                        Exact Conditional Analysis

                         Conditional Exact Tests

                                                --- p-Value ---
          Effect      Test          Statistic    Exact       Mid

          Intercept   Score            3.0000   0.1460    0.1191
                      Probability      0.0537   0.1460    0.1191
          drug        Score            0.4808   0.7278    0.6155
                      Probability      0.2247   0.7278    0.6155
```

```
                        Exact Parameter Estimates

                                    95% Confidence
        Parameter    Estimate           Limits           p-Value

        Intercept     -1.0986     -2.8465      0.2894     0.1460
        drug           0.5091     -1.0858      2.4087     0.7422
```

The SAS output for asymptotic and exact inference when considering the Linear Dependency data set is presented below.

Linear Dependency

	Mutation=NO	**Mutation=YES**	**Total**
Drug=NON-EXPOSURE	**0**	**0**	**0**
Drug=EXPOSURE	**37**	**14**	**51**
Total	**37**	**14**	**51**

```
                          The LOGISTIC Procedure

                            Model Information

                 Data Set                      WORK.TEST
                 Response Variable             mutation
                 Number of Response Levels     2
                 Number of Observations        51
                 Model                         binary logit
                 Optimization Technique        Fisher's scoring


                             Response Profile

                       Ordered                        Total
                         Value     mutation       Frequency

                             1            1              14
                             2            0              37

                       Probability modeled is mutation=1.


                         Model Convergence Status

                 Convergence criterion (GCONV=1E-8) satisfied.


NOTE: The following parameters have been set to 0, since the variables are a linear combination
of other variables as shown.


                            drug =  Intercept
                 Analysis of Maximum Likelihood Estimates

                                       Standard      Wald
                  Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

                  Intercept     1     -0.9719      0.3138        9.5933        0.0020
                  drug          0           0           .         .             .



                             Wald Confidence Interval for Parameters

                        Parameter     Estimate     95% Confidence Limits

                        Intercept      -0.9719      -1.5869      -0.3569
```

```
There was no SAS Output for Exact Conditional Analysis.
```

Results

Complete Separation Data Set:

- Unconditional maximum likelihood estimation (asymptotic inference)

The SAS output provided the following information:

```
Model Convergence Status

      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.
```

The standard errors of the point estimates for the intercept (se= 35.4872) and drug (se= 40.2338) were large compared to the point estimates for the intercept (β = 9.7773) and drug (β = -19.2725). This is typically seen when the maximum likelihood parameter estimates do not converge during the modeling procedure. In addition, one sees that the p-values were both non-significant (p>0.05) for the intercept (p=0.7829) and drug (p=0.6319).

- Exact conditional estimation (exact inference)

The SAS output provided the following information:

A median unbiased estimate of the intercept (MU β = 2.9807, 95% exact CI [1.1991, ∞]) and drug (MU β = -6.4343, 95% exact CI [- ∞ , -4.1539]) are provided. One also sees that the exact p-values were both significant (intercept, p=0.0001; drug, p<0.0001).

Quasi-Complete Separation Data Set:

- Unconditional maximum likelihood estimation (asymptotic inference)

The SAS output provided the following information:

```
Model Convergence Status

  Quasi-complete 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.
```

The standard errors of the point estimates for the intercept (se= 246.0) and drug (se= 246.0) were very large compared to the point estimates for the intercept (β = -13.4954) and drug (β = 12.9155). This is typically seen when the maximum likelihood parameter estimates do not converge during the modeling procedure. In addition, one sees that the p-values were both non-significant (p>0.05) for the intercept (p=0.9562) and drug (p=0.9581).

- Exact conditional estimation (exact inference)

The SAS output provided the following information:

A median unbiased estimate of the intercept (MU β = -2.8224, 95% exact CI [-∞, -1.0219]) and drug (MU β = 2.1708, 95% exact CI [0.2484, ∞]) are provided. One also sees that the exact p-values were both significant (intercept, p=0.0005; drug, p<0.0233).

Overlap Data Set:

- Unconditional maximum likelihood estimation (asymptotic inference)

The SAS output provided the following information:

```
Model Convergence Status

  Convergence criterion (GCONV=1E-8) satisfied.
```

The standard errors of the point estimates for the intercept (se= 0.6666) and drug (se= 0.7455) were of reasonable size compared to the point estimates for the intercept (β = -1.0984) and drug (β = 0.5186). This is typically seen when the maximum likelihood parameter estimates does converge during the modeling procedure. In addition, one sees that the p-values were both non-significant (p>0.05) for the intercept (p=0.0994) and drug (p=0.4867).

- Exact conditional estimation (exact inference)

The SAS output provided the following information:

An exact estimate of the intercept (β = -1.0986, 95% exact CI [-2.8465, 0.2894]) and drug (β = 0.5091, 95% exact CI [-1.0858, 2.4087]) are provided. One also sees that the exact p-values were both non-significant (intercept, p=0.1460; drug, p=0.7422).

Linear Dependency Data Set:

- Unconditional maximum likelihood estimation (asymptotic inference)

The SAS output provided the following information:

```
Model Convergence Status

    Convergence criterion (GCONV=1E-8) satisfied.


NOTE: The following parameters have been set to 0, since the variables are a linear combination
of other variables as shown.


     drug =  Intercept
```

The standard error of the point estimate for the intercept (se= 0.3138) was of reasonable size compared to the point estimates for the intercept (β = -0.9719). This is typically seen when the maximum likelihood parameter estimates does converge during the modeling procedure. The point estimate and the standard error of the point estimate for drug were not computed because of the linear dependency between the two parameters. The p-value for the intercept was significant (p=0.0020).

- Exact conditional estimation (exact inference)

Because of the linear dependency, no SAS output was generated for the exact conditional analysis.

Take Home Points

- The Maximum Likelihood Estimates **do not exist** when you have a data set with **complete separation**.

- The Maximum Likelihood Estimates **may not exist** when you have a data set with **quasi-complete separation**.

- The Maximum Likelihood Estimates **do exist** when you have a data set with **overlap**.

- The Maximum Likelihood Estimates **do exist** when you have a data set with **linear dependency**.

- Exact Conditional Logistic Regression Analysis **can be more informative** than Unconditional Maximum Likelihood Logistic Regression Analysis.

References

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

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

3. Hirji, K.F., Mehta, C.R., and Patel, N.R. (1987), Computing Distributions for Exact Logistic Regression, **Journal of the American Statistical Association**, 82, pp 1110 - 1117.

4. Hirji, K.F., Tsiatis, A.A., and Mehta, C.R. (1989), Median Unbiased Estimation for Binary Data, **American Statistician**, 43, pp 7 - 11.

5. Hirji, K.F. (1992), Computing Exact Distributions for Polytomous Response Data, **Journal of the American Statistical Association**, 87, pp 487 - 492.

6. Kleinbaum, D.G. (1994). **Logistic Regression: A Self-learning Text**. Springer-Verlag, NY, NY. pp 104-119.

7. Mehta, C.R., Patel, N. and Senchaudhuri, P. (1992), Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data, **Journal of Computational and Graphical Statistics**, 1, pp 21 - 40.

8. Mehta, C.R. and Patel, N.R. (1995), Exact Logistic Regression: Theory and Examples, **Statistics in Medicine**, 14, pp 2143 - 2160.

9. Santner TJ and Duffy ED (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, pp 755-758.

10. So Y (1995). A Tutorial on Logistic Regression. **SUGI Proceedings**. http://support.sas.com/techsup/technote/ts450.pdf