

Paper 4654-2020

Rare Events or Non-Convergence with a Binary Outcome? The Power of Firth Regression in PROC LOGISTIC

Patrick Karabon, Oakland University William Beaumont School of Medicine

ABSTRACT

Rare Events and separation are both common analytical challenges encountered when working with a binary variable. Problems with convergence of a logistic regression model **due to complete separation is a particular challenge. Firth's Penalized Likelihood** is a simplistic solution that can mitigate the bias caused by rare events in a data set. Called by the FIRTH option in PROC LOGISTIC, this method will even converge when there is complete separation in a dataset and traditional Maximum Likelihood (ML) logistic regression cannot be run. The implementation of the Firth method is straightforward in SAS® and has **advantages as compared to other potential methods, including Fisher's Exact test, traditional ML logistic regression, and Exact logistic regression.**

This paper briefly introduces the Firth method and discusses the advantages of this method compared to other methods. In addition to the introduction, multiple applications of data **will be used to show SAS® Users when Firth's Penalized Likelihood method might be a good analytic strategy. The applications also will show how to apply Firth's method and provide comparisons between Firth's method and other methods.**

INTRODUCTION

Small sample bias due to either a small data set or a rare outcome may create challenges when analyzing a binary outcome variable using traditional Maximum Likelihood (ML) logistic regression. A phenomenon known as complete separation leads to the **non-convergence of traditional ML logistic regression estimates. Firth's Penalized Likelihood** is a solution used to minimize the analytical bias caused by small samples, rare events, and complete separation.

First introduced by David Firth, the Firth regression originally was a solution to mitigate small sample bias sometimes found in traditional ML logistic regression (Firth, 1993). Subsequently, this method was shown as an effective tool in situations where complete separation within the data does not allow for traditional ML logistic regression estimates to converge (Heinze & Schemper, 2002). The Firth regression falls under the larger umbrellas of penalized likelihood regression techniques, which also include the Least Absolute Shrinkage and Selection Operator (LASSO) regression, the Adaptive LASSO, and the Elastic Net (Gunes, 2015).

The theoretical basis behind **Firth's** regression is that a penalty term is placed on the standard ML function used to generate parameter estimates and standard errors of a logistic regression model. Since the penalty term converges towards 0 as the sample size goes to an infinite number of observation, Firth regression is ideal for small sample bias (Firth, 1993).

This paper will focus on the **real-world application of Firth's Penalized Likelihood** regression to a variety of data sets with complete separation or data sets that have a high potential for small sample bias. For more detailed information on the theory behind Firth regression, the Recommended Reading section contains the original papers on this method.

WHAT IS COMPLETE SEPARATION?

Complete separation, or perfect separation, is the situation where one covariate/explanatory variable always or never occurs with the event of interest/outcome variable. Complete separation tends to happen more often when an event of interest is a rare event; however, complete separation also can occur with non-rare events and even in very large data sets.

As a healthcare-related example, when examining cancer registry data for prostate cancer diagnoses in a population, one would expect that based on biological sex at birth, all of the diagnoses are in males and none are in females. Therefore, it can be said that biological sex **at birth and a prostate cancer diagnosis are complete separable**. In the homeowners' insurance industry, when examining the claims for last year, there were no hurricane damage claims in the state of Michigan. Therefore, it can be said that living in Michigan and submitting an insurance claim for hurricane damage also are completely separable. Finally, as applied to genetic data, suppose there is a specific gene that always accompanies a disease; therefore, the gene and the disease are completely separable (Wang, 2014).

While there are some non-statistical strategies that are recommended to reduce the risk of having complete separation in a data set (Gim & Ko, 2017), elimination of complete separation is not possible in many scenarios and; therefore, a statistical solution such as Firth regression is needed.

Complete Separation or Multicollinearity?

While complete separation is about the relationship between a covariate/explanatory variable and a binary outcome variable, multicollinearity deals with a linear relationship between two covariates/explanatory variables. In the situation of multicollinearity, PROC LOGISTIC will not print parameter estimates for covariate/explanatory variables that are collinear, which makes multicollinearity easy to see in SAS®. However, with complete separation, PROC LOGISTIC still generates a parameter estimate (with warning messages), even though that estimate is very large or very small and is accompanied with an extremely wide confidence interval (Zeng & Zeng, 2019).

In some cases, a User may run into simultaneous problems with multicollinearity and complete separation. There have been additional regression models proposed to handle situations where both problems occur in tandem; however, this falls outside of the scope of this paper (Shen & Gao, 2008).

FIRTH REGRESSION COMPARED TO OTHER METHODS

There are several advantages that make Firth regression an attractive option compared to alternative modeling options, including Fisher's Exact Test, traditional ML logistic regression, and Exact logistic regression.

The most appealing advantage is that the output for Firth regression is almost identical to PROC LOGISTIC output for the standard ML logistic regression. The interpretation of the findings from Firth regression is straightforward for any User familiar with regular logistic regression interpretation.

While there is potential appeal to the Exact methods, there are several disadvantages of the Exact methods. Exact methods are not feasible when a continuously measured covariate/explanatory variable is entered into a model. While categorizing a continuous explanatory variable is possible, it may not be the most desirable option in many circumstances. In addition, the confidence intervals for Exact Logistic regression are very wide, particularly when complete separation is present. When there is complete separation, Exact Confidence Intervals contain either a 0 lower limit or an infinity upper limit depending on the directionality of the complete separation (Heinze, 2006). Finally, Exact methods are very computationally intensive and require long run times. Since even large data sets

(millions have observations) may have complete separation, Exact methods may require more memory than is available from time to time.

In prior simulation studies, Firth regression was successfully applied to a variety of data sets. Type I Error Rates were consistently low and convergence was almost never an issue. Firth regression estimates were considered "highly efficient" in the presence of rare events or complete separation (Heinze, 2006).

IMPLEMENTING FIRTH REGRESSION IN SAS/STAT®

Calling Firth regression is very simple in SAS/STAT®. The method is called by adding the FIRTH option into the MODEL statement of PROC LOGISTIC.

Note that there also is a FIRTH option in the MODEL statement of PROC PHREG. This paper will not cover any examples using PROC PHREG, but Firth Regression can also be applied in the context of survival/duration analysis.

ILLUSTRATIVE EXAMPLES

With the exception of Example #4, the data for the first three examples has been simulated in order to provide the best possible examples for Firth regression.

EXAMPLE #1: FIRTH REGRESSION AND COMPLETE SEPARATION

In this first example, a comparison of two surgical procedures (binary variable PROCEDURE) is being analyzed to examine if an association exists between the two procedures and subsequent development of a specific complication (binary outcome COMPLICATION). This is a retrospective pilot study on 30 cases using the new procedure and 200 cases using the old procedure.

The following code calls PROC FREQ to obtain a 2 x 2 contingency table with column percentages and also requests a Chi-Square test, a Fisher's Exact test, and Relative Risk measures (Relative Risk and Odds Ratio):

```
PROC FREQ DATA = SPARSE;
    TABLE COMPLICATION*PROCEDURE / NOPERCENT NOROW CHISQ FISHER RELRISK;
RUN;
```

A 2 x 2 contingency table to evaluate the association between PROCEDURE and COMPLICATION is shown below:

COMPLICATION	PROCEDURE		
	New	Old	Total
Yes	0	9	9
	0.00	4.50	
No	30	191	221
	100.00	95.50	
Total	30	200	230

Table 1: 2 x 2 Contingency Table

As shown above in Table 1, 9 out of 200 (4.50%) cases with the Old Procedure had a Complication while none of the 30 (0.00%) cases with the New Procedure had a Complication. Therefore, Complete Separation exists between the outcome variable PROCEDURE and the covariate/explanatory variable COMPLICATION.

First, we are going to look at the corresponding output in PROC FREQ, beginning with the summarized Chi-Square output as shown below in Table 2:

Statistic	DF	Value	Prob
Chi-Square	1	1.4050	0.2359
...
WARNING: 25% of the cells may have expected counts less than 5. Chi-Square may not be a valid test.			

Table 2: Chi-Square Test Results

While output is printed for the Chi-Square statistic, the output comes along with a warning message, which states that Chi-Square may not be a valid test due to the assumption of adequate expected counts not being met. Since the assumption of expected counts is not met, a suitable alternative is the Fisher’s Exact test, which also is requested using PROC FREQ as shown below:

Cell (1, 1) Frequency (F)	30
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	0.2775
Table Probability (P)	0.2775
Two-sided Pr <= P	0.6098

Table 3: Fisher's Exact Test Results

The Fisher’s Exact test shows that there is not enough evidence to show a statistically significant association in the rate of complications between the two procedures (Two-Sided P = 0.6098). However, in this case, our PI/Client wants an effect size measurement to go along with the P-Value. Since this study is retrospective, then the Odds Ratio (OR) is a good measure of effect size. In PROC FREQ, Odds Ratios can be requested through the RELRISK option:

Statistic	Value	95% Confidence Limits	
Relative Risk (Column 2)	0.8643	0.8203	0.9106
One or more statistics not computed – zero cell.			

Table 4: Odds Ratio and Relative Risk Results

Due to complete separation, an Odds Ratio estimate is not possible. The RELRISK output comes along with a warning message that states “One or more statistics not computed – zero cell” and warns us of the complete separation.

Since it is not possible to get an Odds Ratio estimate from PROC FREQ, PROC LOGISTIC is an alternative procedure because it also produces Odds Ratio estimates as part of its standard output. The following code will produce the Odds Ratio estimate for traditional Maximum Likelihood logistic regression:

```
PROC LOGISTIC DATA = SPARSE;
  CLASS COMPLICATION(REF = "No") PROCEDURE(REF = "Old") / PARAM = GLM;
  MODEL COMPLICATION = PROCEDURE;
RUN;
```

However, even before looking the Output, there are two separate warning messages printed in the log as shown in Output 1:

```
WARNING: There is a possibility a quasi-complete separation of data points.
The maximum likelihood estimate may not exist.

WARNING: The LOGISTIC procedure continues in spite of the above warning.
Results are shown based on the last maximum likelihood iteration. Validity
of the model fit is questionable.
```

Output 1: Warning Messages in Log

In addition, the Output Window has two separate Warning messages written in it as well:

```
WARNING: The maximum likelihood estimate may not exist.

WARNING: The LOGISTIC procedure continues in spite of the above warning.
Results are based on the last maximum likelihood iteration. Validity of the
model fit is questionable.
```

Output 2: Warning Messages in PROC LOGISTIC Output Window

Since there are four separate warnings that valid estimates may not exist, we should not continue with traditional Maximum Likelihood logistic regression. Exact logistic is historically a viable alternative to obtain an Odds Ratio. Exact logistic regression is called by adding the EXACT statement into PROC LOGISTIC. Either the PARAM = GLM or PARAM = REF model parameterization options must be specified in the CLASS statement to Exact Logistic to run. The following code runs the Exact logistic regression:

```
PROC LOGISTIC DATA = SPARSE;
  CLASS COMPLICATION(REF = "No") PROCEDURE(REF = "Old") / PARAM = GLM;
  MODEL COMPLICATION = PROCEDURE;
  EXACT PROCEDURE / ESTIMATE = ODDS;
RUN;
```

From the above code, the Exact Odds Ratio output is as follows in Table 5:

Parameter		Estimate		95% Confidence Limits	P-Value	
PROCEDURE	New	0.518	*	0	2.641	0.2775

Table 5: Exact Logistic Regression Results

Note that from Table 5 above, the 95% confidence limits have a lower bound of 0, which implies a log odds of negative infinity. Therefore, there is no finite confidence interval using Exact logistic regression. Firth regression might help with this issue of no finite confidence interval. The Firth regression, called by the FIRTH option in the model statement, is run using the following code:

```
PROC LOGISTIC DATA = SPARSE;
  CLASS COMPLICATION(REF = "No") PROCEDURE(REF = "Old") / PARAM = GLM;
  MODEL COMPLICATION = PROCEDURE / FIRTH;
RUN;
```

The Parameter Estimates output of Firth regression is as follows in Table 6:

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0036	0.3332	81.2481	< 0.0001
PROCEDURE	New	1	-1.1078	1.4875	0.5547	0.4564
PROCEDURE	Old	0	0	.	.	.

Table 6: Firth Regression Parameter Estimates

In addition, the Odds Ratio estimates are as follows:

Effect	Point Estimate	95% Wald Confidence Limits	
		Lower	Upper
PROCEDURE New vs Old	0.330	0.018	6.096

Table 7: Firth Regression Odds Ratio Estimates

As is shown above, there is finally a finite effect size estimate in addition to finite 95% confidence limits. Even with Firth regression, we also conclude that there is not enough evidence that the New Procedure has significantly higher Complications ($P = 0.4564$). This first example demonstrates that even with complete separation, Firth regression provides an Odds Ratio with a finite 95% confidence interval.

EXAMPLE #2: WORKING WITH A CONTINUOUS COVARIATE

In Example #2, the same dataset as Example #1 is used; however, we want to examine the effect of age (a continuous covariate/explanatory variable) on the binary outcome Complication. While we could categorize age into groups or just run a Two Samples Independent T-Test, age will remain a continuously measured variable and logistic regression is used for the sake of this example. The breakdown of age by those with and without complications is as follows:

Complication	N	Mean	Standard Deviation
Yes	9	43.44	5.81
No	221	44.61	6.12

Table 8: Descriptive Statistics for Age

Complication is not completely separated here, rather, it is a rare event with a 3.9% rate of complications (9 complications out of 230 cases). It does not appear that there is much of an association between age and complication by just looking at the above numbers in Table 8.

Since age is a continuously measured variable, Exact methods (Fisher's Exact Test and Exact logistic regression) are not possible; therefore, we will compare the results of traditional Maximum Likelihood logistic regression and Firth regression with the following code for both, respectively:

```

/* Maximum Likelihood Logistic Regression */
PROC LOGISTIC DATA = SPARSE;
  CLASS COMPLICATION(REF = "No") / PARAM = GLM;
  MODEL COMPLICATION = AGE;
RUN;

/* Firth Regression */
PROC LOGISTIC DATA = SPARSE;
  CLASS COMPLICATION(REF = "No") / PARAM = GLM;
  MODEL COMPLICATION = AGE / FIRTH;
RUN;

```

Some amended output to compare the parameter estimates for both Maximum Likelihood and Firth regression are as follows:

Model	Estimate	Standard Error	Odds Ratio	95% Wald Confidence Limits		Wald Chi-Square	P-Value
				Lower	Upper		
Logistic	-0.0345	0.0615	0.966	0.856	1.090	0.3159	0.5741
Firth	-0.0246	0.0572	0.976	0.872	1.091	0.1853	0.6669

Table 9: Comparison of Model Estimates

There are a few things of interest in Table 9. The standard error for the Firth regression is slightly smaller than the standard error for Maximum Likelihood, which leads to a slightly narrower 95% confidence interval for Firth regression. Also, the Odds Ratio is slightly close to 1 (the null hypothesis) for Firth regression than for traditional Maximum Likelihood logistic regression.

EXAMPLE #3: LARGE DIFFERENCE IN TWO GROUPS WITH A RARE EVENT

In this third example, we are using a different dataset, which has a rare event with no complete separation of the data. The binary categorical covariate/explanatory variable is GROUP while the binary outcome is EVENT. A 2 x 2 contingency table between the EVENT and GROUP variable is as follows:

EVENT	GROUP		Total
	1	2	
Yes	4	15	19
	26.67	7.14	
No	11	195	206
	73.33	92.86	
Total	15	210	225

Table 10: 2 x 2 Contingency Table

We want to compare the parameter estimates, 95% confidence intervals, and P-Values for Maximum Likelihood logistic regression, Exact logistic regression, and Firth regression. The following code shows how to run all three models:

```

/* Maximum Likelihood Logistic Regression */
PROC LOGISTIC DATA = SPREAD;
  CLASS EVENT(REF = "No") GROUP(REF = "2") / PARAM = GLM;
  MODEL EVENT = GROUP;
RUN;

/* Exact Logistic Regression */
PROC LOGISTIC DATA = SPREAD;
  CLASS EVENT(REF = "No") GROUP(REF = "2") / PARAM = GLM;
  MODEL EVENT = GROUP;
  EXACT GROUP / ESTIMATE = ODDS;
RUN;

/* Firth Regression */
PROC LOGISTIC DATA = SPREAD;
  CLASS EVENT(REF = "No") GROUP(REF = "2") / PARAM = GLM;
  MODEL EVENT = GROUP / FIRTH;
RUN;

```

The subsequent amended output of this code is shown below in table form:

Model	Estimate	Standard Error	Odds Ratio	95% Wald Confidence Limits		Wald Chi-Square	P-Value
				Lower	Upper		
Logistic	1.5533	0.6424	4.727	1.342	16.651	5.8465	0.0156
Exact	.	.	4.673	0.969	18.436	.	0.0549
Firth	1.5963	0.6323	4.935	1.429	17.040	6.3744	0.0116

Table 11: Comparison of Regression Estimates

There are a few interesting findings from the above Table. While the associations for traditional Maximum Likelihood logistic regression and Firth regression are significant, the results for Exact logistic regression are not significant. As compared to traditional Maximum Likelihood Logistic regression, Firth regression has a slightly higher Odds Ratio, but a

slightly smaller standard error and thus a narrower confidence interval. This leads to a slightly more significant P-Value in Firth regression; however, both P-values are very similar.

EXAMPLE #4: COMPARISON OF RUN TIMES

For advanced Users of SAS®, one of the well-known limitations of Exact methods (Fisher's Exact Test and Exact Logistic regression) is that the run times can be lengthy and they can use lots of memory. However, less is known whether Firth regression takes much more time and uses much more memory than traditional Maximum Likelihood logistic regression.

In this example using real-world data, the data is being sequentially collected on over 2 million observations. As we collect each additional 100,000 observations, we are going to check the CPU run times of the various models in SAS/STAT® and compare them between traditional Maximum Likelihood logistic regression, Exact logistic regression, and Firth regression.

Figure 1 below compares the CPU times, in seconds, between traditional Maximum Likelihood logistic regression and Firth regression. Exact logistic regression is missing because Exact regression began to have memory size problems at 20,000 observations and this example was run on a powerful and robust system. Depending on the model specification, memsize can be increased; however, memsize is a finite parameter for everyone so eventually there will be no space left to run the Exact method.

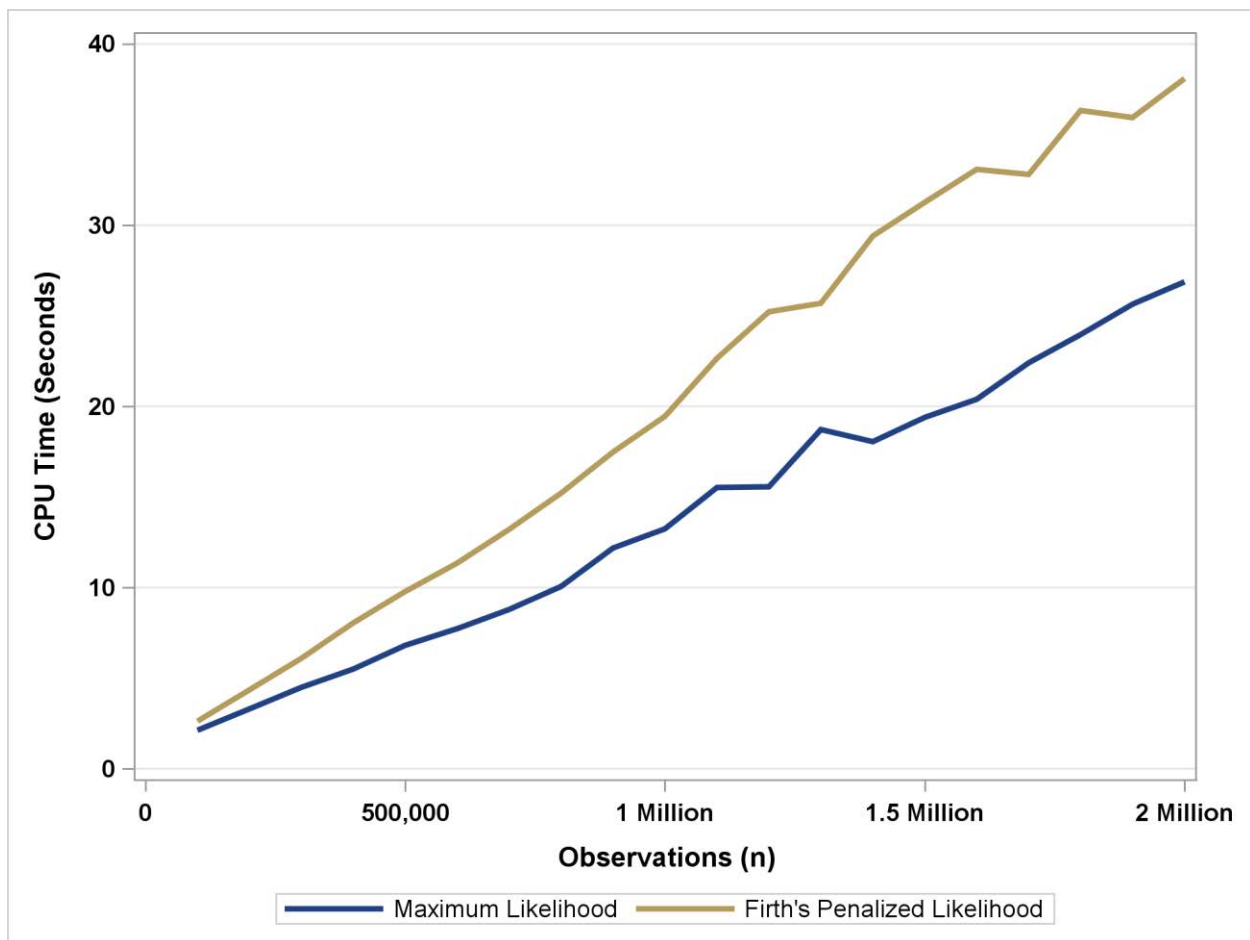


Figure 1: Comparison of Runtimes

While Firth regression has longer run times than traditional Maximum Likelihood logistic regression, all run times were less than 40 seconds of CPU time even for a very large

dataset of 2 million observations. While there are longer run times for Firth regression, the difference in run times is not too large as to cause an inconvenience in running the Firth regression with a large dataset.

Would Firth regression be needed with a large dataset of over 1 million observations? While small sample bias is not a concern with these very large datasets, it is possible that complete separation can still occur; therefore, Firth regression may be necessary even with very large datasets.

CONCLUSION

In conclusion, Firth regression is a good alternative to traditional Maximum Likelihood logistic regression when there are rare events or when complete separation exists. Firth regression produces finite parameter estimates even when complete separation exists. Run times and memory required to run Firth regression do not impose a burden to those using this method. Finally, Firth regression is very easy to run and interpret for anybody familiar with PROC LOGISTIC.

REFERENCES

- Firth, D. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika*, 80: 27-38.
- Gim, T. H. T. and Ko J. 2018. "Maxim Likelihood and Firth Logistic Regression of the Pedestrian Route Choice." *International Regional Science Review*, 40: 616-637.
- Gunes, F. 2015. "Penalized Regression Methods for Linear Models in SAS/STAT®." *Proceedings of the SAS Global Forum 2015*. Dallas, Texas. Available at: http://support.sas.com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels.pdf
- Heinze, G. 2006. "A comparative investigation of methods for logistic regression with separated or nearly separated data." *Statistics in Medicine*, 25: 4216-4226.
- Heinze, G. and Schemper, M. 2002. "A solution to the problem of separation in logistic regression." *Statistics in Medicine*, 21: 2409-2419.
- Shen, J. and Gao, S. 2008. "A solution to Separation and Multicollinearity in Multiple Logistic Regression." *Journal of Data Science*, 6: 515-531.
- Wang, X. 2014. "Firth logistic regression for rare variant association tests." *Frontiers in Genetics*, 5: 1-2.
- Zeng, G. and Zeng, E. 2019. "On the relationship between multicollinearity and separation in logistic regression." *Communications in Statistics – Simulation and Computation*.

RECOMMENDED READING

- *SAS/STAT® User's Guide: Example 73.13 Firth's Penalized Likelihood Compared with Other Approaches*. Available at: https://documentation.sas.com/?docsetId=statug&docsetVersion=14.2&docsetTarget=statug_logistic_examples15.htm&locale=en.

CONTACT INFORMATION

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

Patrick Karabon, MS
Oakland University William Beaumont School of Medicine
pkarabon@oakland.edu
www.oakland.edu/medicine