Paper ST-07

Tailoring Logistic Regression Model Analyses with the ODDSRATIO Statement in PROC LOGISTIC

Taylor Lewis, University of Maryland, College Park, MD

ABSTRACT

Binary logistic regression is typically preferred when modeling a dichotomous outcome variable. Interpretation can be tricky, however, since parameter estimates of the model are given in terms of log-odds. Exponentiating the log-odds returns an odds ratio, which is somewhat easier to handle. By default, PROC LOGISTIC will output a series of odds ratios for all predictor variables not involved in any interactions. The new ODDSRATIO statement offers the flexibility to tailor odds ratios per the analyst's desired comparisons, even when interactions are specified. In addition to discussing odds ratios for categorical variables, this paper illustrates how the UNITS statement can facilitate customized odds ratios for continuous explanatory variables.

INTRODUCTION

Unlike linear regression where the outcome variable is measured on a continuous scale, logistic regression is appropriate for modeling a binary variable, one in which an *event* or *non-event* occurs. Applying linear regression techniques to a dichotomous (0 or 1) outcome violates key theoretical assumptions—for instance, values of the outcome are finite and bounded, with the error term distributed binomially, not normally. Employing the *logit*, or *log-odds*, transformation alleviates these issues. There are several SAS® procedures capable of fitting logistic regression models, but this paper discusses PROC LOGISTIC and highlights the ODDSRATIO statement new to SAS/STAT® in Version 9.2.

By analyzing a hypothetical data set motivated by the National Immunization Survey,[1] the interpretation of logistic regression coefficients and odds ratios are discussed alongside PROC LOGISTIC syntax and output. We assume the SAS data set VACCINATIONS contains 1,009 observations from a survey sponsored by a state public health agency to estimate the proportion of children aged 19 to 36 months who have been vaccinated for a particular virus. The key outcome variable is UTD, taking on a value of 1 if the child is up-to-date with respect to the vaccination and 0 otherwise. On top of the obvious state-level, univariate analysis, we may also wish to model vaccination likelihood using other demographic and socioeconomic variables collected from the survey. Example variables and considerations relevant to logistic regression model building are introduced in the next section.

INTERPRETING THE LOGISTIC REGRESSION MODEL AND ODDS RATIOS

In linear regression with a single continuous variable *X*, we might suppose a linear relationship such as $E(Y \mid X = x) = \beta_0 + \beta_1 x$, interpreting β_1 as the expected change in *Y* given a one-unit increase in *X*. In logistic regression, the logit transformation yields $L(Event \mid X = x) = \log\left(\frac{\Pr(Event \mid x)}{1 - \Pr(Event \mid x)}\right) = \beta_0 + \beta_1 x$, implying a one-unit increase in *X* brings about a β_1 change in the logit (the log-odds).

Returning to the vaccination data, the following syntax allows us to fit a logistic regression model relating child age in months (AGE) to the likelihood of being up-to-date:

```
proc logistic data=vaccinations;
  model UTD (event='1') = age;
run;
```

By default, SAS assumes the first value from the sort order of the response variable is the event. In the present case, this corresponds to a child having *not* been vaccinated (UTD=0). We can use the `EVENT=` option in parentheses immediately following the outcome variable in the MODEL statement to assign the desired code

[1] www.cdc.gov/nis

(UTD=1). An alternative method is to specify the keyword DESCENDING in the PROC statement, which reverses the sort order prior in the event assignment rule.

PROC LOGISTIC generates a lot of output. Only relevant output is presented in this paper as it fits with the discussion at hand. Estimated model parameters appear in the Analysis of Maximum Likelihood Estimates section:

```
               Analysis of Maximum Likelihood Estimates

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

Intercept      1     -0.8265      0.4103        4.0575        0.0440
age            1      0.0835      0.0157       28.3183        <.0001
```

The model is estimated to be $L(Vaccination \mid age) = \hat{\beta}_0 + \hat{\beta}_1 age = -0.8265 + 0.0835 * age$. After a little algebra, the predicted probability of vaccination can be extracted from the logit function by $\Pr(Vaccination \mid age) = \dfrac{e^{-0.8265+0.0835*age}}{1+e^{-0.8265+0.0835*age}} = \left(1+e^{-(-0.8265+0.0835*age)}\right)^{-1}$. To provide a few numerical examples, a 20-month-old child has a 0.699 probability of vaccination, while the probability for a 30-month-old child is 0.843. The vaccination probability thus increases with age, a sensible result.

Since $L(Vaccination \mid age = x+1) - L(Vaccination \mid age = x) = \left(\hat{\beta}_0 + \hat{\beta}_1(x+1)\right) - \left(\hat{\beta}_0 + \hat{\beta}_1(x)\right) = \hat{\beta}_1$, we can interpret $\hat{\beta}_1$ as the estimated change in the log-odds given a one-unit increase in age. If we exponentiate, $e^{\hat{\beta}_1}$ is interpreted as the estimated *odds ratio* between two children one-month apart in age. PROC LOGISTIC will exponentiate all model parameters when the `EXPB` option is given after the slash in the MODEL statement:

```
proc logistic data=vaccinations;
  model UTD (event='1') = age / expb;
run;
```

```
                    Analysis of Maximum Likelihood Estimates

                                Standard          Wald
Parameter     DF    Estimate       Error    Chi-Square    Pr > ChiSq    Exp(Est)

Intercept      1     -0.8265      0.4103        4.0575        0.0440       0.438
age            1      0.0835      0.0157       28.3183        <.0001       1.087
```

We note $e^{\hat{\beta}_1} = e^{0.0835} = 1.087$, which matches what is output under the Exp(Est) column. Colloquially, we often conclude a child one-month older is 8.7% more likely to be vaccinated. In truth, such an interpretation is more apt for the relative risk statistic—the ratio of the two predicted probabilities. The tendency to treat the odds ratio as if were the relative risk may be derived from the analysis of rare events, in which case the two statistics approximate each other (p. 50 of Hosmer and Lemeshow, 2000).

For each variable in the MODEL statement, odds ratios are also printed to the Odds Ratio Estimates section of the output:

```
             Odds Ratio Estimates

               Point          95% Wald
Effect      Estimate      Confidence Limits

age            1.087       1.054       1.121
```

Odds ratios are asymmetric, ranging from 0 to ∞, and a value near 1indicates no difference. A “significant” odds ratio is one whose confidence interval does not include 1. For the data at hand, even a one-month age difference results in a significant increase in the odds of vaccination.

The simple logistic regression model illustrated thus far has included a sole continuous covariate. When an explanatory variable is categorical with, say, *k* distinct values, one must create *k* – 1 *indicator variables* (SAS calls them *design variables*) which take on a value of 1 if the observation belongs to the k^{th} group and 0 otherwise. The process of creating these indicator variables is known as *parameterization*. The group which has no indicator variable (i.e., no model parameter) is called the *reference* group, so named because the parameters for all other *k* – 1 variables estimate the change in the log-odds as compared to that category.

Assume we want to add the effect of race into the model. This would allow us to analyze the odds ratios for age differences while simultaneously accounting for race, and vice versa. In the VACCINATIONS data set, the variable RACE denotes a child's minority status, with whites coded as 1 and all other races coded as 2. We can run the following augmented syntax:

```
proc logistic data=vaccinations;
  class race / param=ref;
  model UTD (event='1') = age race / expb;
run;
```

SAS automatically creates the *k* – 1 indicator variables for all categorical effect specified in the CLASS statement. The PARAM=REF option after the forward slash overrides the default *effect* parameterization, which is not as easily interpretable in this author's opinion—for more discussion, see Lewis (2007). The Class Level Information portion of the output summarizes how SAS parameterized the variable RACE:

```
   Class Level Information

                     Design
Class     Value     Variables

race      1                1
          2                0
```

PROC LOGISTIC sorts the given class variable and assigns the last value to be the reference group. To modify this election, simply provide the reference group code in the REF=' ' option within parentheses after RACE has been listed in the CLASS statement.

Now the Analysis of Maximum Likelihood Estimates section contains an additional parameter to represent the effect of a child being white versus minority. The estimated odds ratio for age has changed slightly after incorporating race as a covariate into the model. Once again, the exponentiated model coefficient matches what is printed in the Odds Ratio Estimates section:

```
                  Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter       DF    Estimate       Error    Chi-Square    Pr > ChiSq    Exp(Est)

Intercept        1     -1.1107      0.4535        5.9982        0.0143       0.329
age              1      0.0842      0.0157       28.6170        <.0001       1.088
race      1      1      0.3166      0.2112        2.2476        0.1338       1.373

             Odds Ratio Estimates

                  Point          95% Wald
Effect         Estimate      Confidence Limits

age               1.088       1.055       1.122
race 1 vs 2       1.373       0.907       2.076
```

Odds ratios are computed for all pairwise differences of the categorical variable, regardless of the parameterization. Since RACE is coded as a simple dichotomy, there is only one comparison, but more comparisons will be appear for variables with three or more distinct categories.

COMPLICATIONS

Odds ratios have been fairly straightforward to glean from PROC LOGISTIC up until this point, but a few complications can arise. For one, odds ratios are not restricted to one-unit differences in a continuous *X* variable. As an example, we could compare the odds ratio of vaccination at 30 months versus 20 months given the same child race by calculating $L(Vaccination \mid age = 30) - L(Vaccination \mid age = 20) = \left(\hat{\beta}_0 + \hat{\beta}_1(30)\right) - \left(\hat{\beta}_0 + \hat{\beta}_1(20)\right) = 10\hat{\beta}_1$ then exponentiating to get $e^{10\hat{\beta}_1} = e^{10*0.0842} = 2.320$. This is possible due to the property $\log\left(\frac{L_1}{L_2}\right) = \log\left(L_1 - L_2\right)$.

Thus, exponentiating the difference between two logits returns an odds ratio.

The UNITS statement in PROC LOGISTIC facilitates this customization:

```
proc logistic data=vaccinations;
  class race / param=ref;
  model UTD (event='1') = age race / expb;
units age=10;
run;
```

```
             Odds Ratios

Effect                Unit     Estimate

age                10.0000        2.320
```

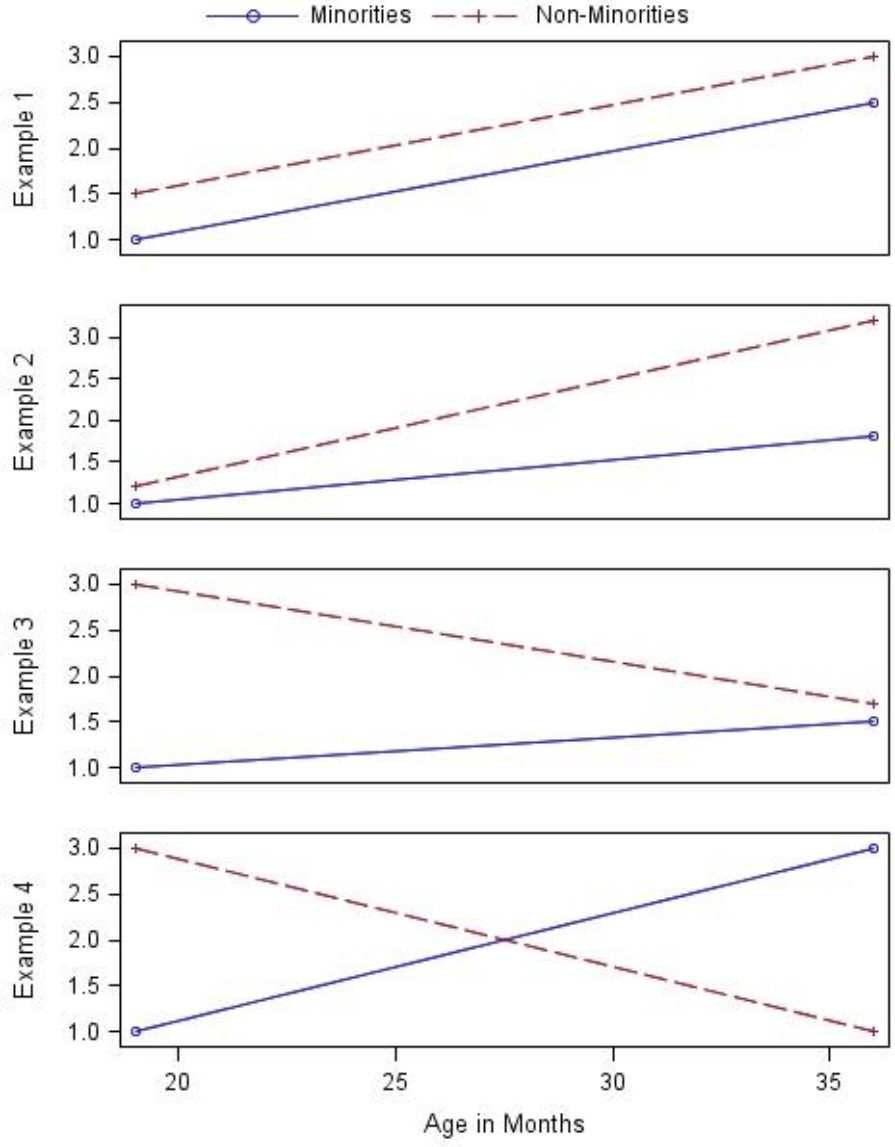
The estimated odds ratio for a ten-unit change in age is output to a new component of the listing, Odds Ratios, which appears after the Odds Ratio Estimates component seen previously. Unfortunately, no confidence interval is given to assess significance. One could estimate the variance by hand using the variance-covariance matrix of the model coefficients (see Section 2.5 of Hosmer and Lemeshow, 2000), but the process can be a little tedious. As will be demonstrated shortly, a few lines of code in SAS' new ODDSRATIO statement is all that is needed to complete the analysis.

Interaction terms in the model can also complicate matters. The Odds Ratio Estimates portion of the output only lists variables not interacting with any other variables in the model. In a way, this protects the analyst because it is not always wise to make marginal odds ratio inferences based on a variable that interacts with another. Figure 1 illustrates this concept.

In the topmost plot (Example 1), age has a linearly increasing effect on the logit for both minorities and non-minorities. The two lines' vertical distance is an estimate of the logit difference between the two racial categories, which appears more or less constant across all ages—that is, the two lines are roughly parallel. This is a situation where there is likely no interaction between the two variables.

Examples 2 through 4 illustrate various types of scenarios implying an interaction exists. The logit difference is not constant across all ages in Example 2: for younger ages, white children are slightly more likely to be vaccinated, but the disparity widens as child age increases. Example 3 shows the opposite trend, where the inequity lessens with age. Example 4 is a case where the expected logit by age clearly hinges on minority status. Interestingly, this is a circumstance where we might find the two variables' respective model coefficients insignificant, yet their interaction highly significant.

Figure 1. Four Hypothetical Logit Relationships of Age and Minority Status on the Likelihood of Vaccinations.



THE ODDSRATIO STATEMENT

The ODDSRATIO statement can be utilized to make inferences on customized odds ratio estimates, even with interaction terms in the underlying model. Suppose we increase the complexity of the vaccination likelihood model by including a term representing the interaction between child age and minority status as well as introducing a 3-category household income level variable INCCAT, where larger values indicate higher income levels. The syntax and corresponding output are as follows:

```
proc logistic data=vaccinations;
  class race inccat / param=ref;
  model UTD (event='1') = age|race @2 inccat / expb;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	0.000621	1.0248	0.0000	0.9995	1.001
age		1	0.0392	0.0373	1.1057	0.2930	1.040
race	1	1	-0.9194	1.1068	0.6901	0.4061	0.399
age*race	1	1	0.0472	0.0410	1.3278	0.2492	1.048
inccat	1	1	0.2319	0.2392	0.9400	0.3323	1.261
inccat	2	1	-0.1716	0.2377	0.5209	0.4704	0.842

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
inccat 1 vs 3	1.261	0.789	2.015
inccat 2 vs 3	0.842	0.529	1.342

A pipe between two or more variables followed by `@N` is shorthand syntax telling SAS to include in the model each variable and all possible N-way interactions. In the example above, `age|race @2` is equivalent to specifying `age race age*race`, and either method bypasses the need to create interaction variables in a separate DATA step.

INCCAT is the only variable appearing in the Odds Ratio Estimates portion of the output. AGE and RACE, because they are involved in an interaction, are excluded. The interaction allows the effect of race to differ according to age. If we believe such a relationship holds, it might be prudent to restrict the racial comparison to a particular age value(s). The ODDSRATIO statement reasons this way:

```
proc logistic data=vaccinations;
  class race inccat / param=ref;
  model UTD (event='1') = age|race @2 inccat / expb;
oddsratio race;
run;
```

Wald Confidence Interval for Odds Ratios

Label	Estimate	95% Confidence Limits	
race 1 vs 2 at age=27.056	1.430	0.940	2.177

Output from the ODDSRATIO statement is sent to the Wald Confidence Interval for Odds Ratio section. From the label, we gather PROC LOGISTIC's default method is to calculate the odds ratio for a categorical variable at the mean level of the continuous variable with which it interacts. Such a value may or may not be meaningful. To request specific ages at which to evaluate the odds ratio difference, we can augment the syntax as follows:

```
proc logistic data=vaccinations;
  class race inccat / param=ref;
  model UTD (event='1') = age|race @2 inccat / expb;
oddsratio race / at(age=20 25 30);
run;
```

Wald Confidence Interval for Odds Ratios

Label	Estimate	95% Confidence Limits	
race 1 vs 2 at age=20	1.025	0.525	2.001
race 1 vs 2 at age=25	1.298	0.840	2.005
race 1 vs 2 at age=30	1.644	0.994	2.718

Now we are given the white versus minority comparisons at three distinct age slices. There is some evidence the odds ratio increases with race, a scenario depicted by Example 2 in Figure 1, although the increase is subtle.

The ODDSRATIO statement can accommodate only one variable at a time, but multiple statements are allowed. If we place AGE in another ODDSRATIO statement without any additional options, the output is extended with the following:

```
Label              Estimate    95% Confidence Limits

age at race=1         1.090       1.054        1.127
age at race=2         1.040       0.967        1.119
```

Hence, for a continuous variable interacting with a categorical variable, a one-unit odds ratio is provided across all *k* distinct categories. To select particular comparisons the same `AT(VAR=)` syntax can be used after the forward slash in the ODDSRATIO statement, except values for which `VAR=` references a categorical variable should be enclosed in quotes. For instance, the output above would be reproduced with the following syntax:

```
proc logistic data=vaccinations;
  class race inccat / param=ref;
  model UTD (event='1') = age|race @2 inccat / expb;
oddsratio age / at(race='1' '2');
run;
```

More complicated model settings can be handled, but the recipe for analysis is similar. Had we changed the model statement above to include a second two-way interaction between AGE and INCCAT, the customized odds ratios for age would need to control for a particular level of INCCAT in addition to RACE. Without doing so, an age odds ratio is computed for all combinations of INCCAT x RACE. Similar logic applies for three-way interactions.

Another useful feature of the ODDSRATIO statement is that, when used in conjunction with the UNITS statement, one can calculate a confidence interval for a customized continuous variable odds ratio. This is true even if the model lacks interactions.

Recall earlier we wanted to estimate the odds ratio for a 10-month age increase under the model which included age and race. The UNITS statement alone gave only the point estimate. By adding the ODDSRATIO statement, the desired confidence interval can be obtained from the Wald Confidence Interval for Odds Ratio section:

```
proc logistic data=vaccinations;
  class race / param=ref;
  model UTD (event='1') = age race / expb;
units age=10;
oddsratio age;
run;
```

```
      Wald Confidence Interval for Odds Ratios

Label                Estimate    95% Confidence Limits

age units=10            2.320       1.705        3.158
```

CONCLUSION

This paper exposited some of the key relationships behind properly interpreting a logistic regression model—namely, the relationship between parameter estimates, logit differences, and odds ratios. Understanding how these statistics relate to one another is vital when customizing analyses based on the model fit via PROC LOGISTIC. By default, PROC LOGISTIC outputs many pertinent odds ratio comparisons, but only for variables not involved in any interactions. For users who wish to customize odds ratios for variable(s) involved in one or more interactions, the ODDSRATIO statement, new to SAS Version 9.2, offers an additional avenue for analysis.

An interaction term is necessary when one variable's effect varies depending upon the level of a second variable. Thus, when making an odds ratio comparison for the first variable, it is wise to control the level of the second with the `AT(var=)` option after the forward slash in the ODDSRATIO statement. This recommendation is particularly poignant if the variable in the ODDSRATIO statement interacts with a continuous variable. In that scenario, PROC LOGISTIC defaults to an odds ratio calculated at its mean, which may be an arbitrary or unimportant value.

The UNITS statement allows users to compute odds ratios for continuous variable odds in increments other than 1. We have seen, however, that only a point estimate is provided. To get a confidence interval, one can use the UNITS statement in combination with the ODDSRATIO statement.

Lastly, the example data in this paper was motivated by the National Immunization Survey (NIS). It should be acknowledged that survey data often contain features such as stratification, clustering, and differential weights which require the use of the SURVEY family of SAS procedures. For instance, the analog to PROC LOGISTIC is PROC SURVEYLOGISTIC. Discussion was purposefully limited to the former procedure, since the ODDSRATIO statement is not yet available in PROC SURVEYLOGISTIC.

REFERENCES

Hosmer, D. and Lemeshow, S. (2000), *Applied Logistic Regression*, Second Edition. New York, NY: Wiley.

Lewis, T. (2007). "PROC LOGISTIC: The Logistics Behind Interpreting Categorical Variable Effects." Proceedings of the 20th Annual Northeast SAS User's Group (NESUG) Conference. Baltimore, MD.

CONTACT INFORMATION

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

Taylor Lewis
Joint Program in Survey Methodology
1218 LeFrak Hall
University of Maryland
College Park, MD 20742
E-mail: tlewis@survey.umd.edu