**Chapter 13.7: Multiple Logistic Regression**

**File: chapt13LogisticRegAct12**

**Will be part of Exam (no HW on this topic)**

**Objectives**

In this lesson, we address multiple logistic regression for modelling dichotomous categorical response using continuous and categorical predictors via SAS implementation.

The specific objectives are as follows:

* Understand odds and odds ratio
* Logit transformation
* Briefly introduce a logistic regression
* Fit a logistic regression model via SAS
* Interpret results for the response via odds ratio of predictor of one group against the ref group.
* Evaluate prediction accuracy of the model

**Basic concepts**

Let the random variable take two values 1 (success or event of interest) and 0 (failure) with probabilities

and

It is easy to see that , taking value between 0 and 1.

Then, the random variable follows a Bernoulli distribution with success probability .

**Odds and odds ratio**

The quantity is called the **odds of success or odds in favor of a success** .

Let be probability of disease among males. Then, and are called odds of disease among males and females, respectively.

The quantity is called odds ratio of disease among males compared to females.

**Log-odds and log-odds ratio**

The quantities and are called log-odds of disease among males and females, respectively. Then, the ratio

is called log-odds ratio of disease among males compared to females.

**Logit function**

The logit of is the function .

**Example 13.7.1**

Suppose that the probability of a certain disease among males in a population is , while the same among females is .

Compute the following:

1. Odds of disease among males
2. Odds of disease among females
3. Odds ratio of disease among males compared to females
4. Odds ratio of disease among females compared to males.
5. Compute log-odds ratio of disease among males compared to females.
6. Compute the logit of 0.75.

Solution

1. The odds of disease among males is:
2. The odds of disease among females is:
3. The odds ratio of disease among males compared to females is:

That is, the odds of having the disease among males is 3.478 times the odds of females. So,

1. The odds ratio of disease among females compared to males is:

That is, the odds of having disease among females is 0.29 times the odds of males.

So,

1. The log-odds ratio of disease among males compared to females is:

(f)

**Logistic regression**

A logistic regression models the ***mean response*** of a dichotomous (or polytomous) variable by a ***set of predictors*** via a ***logit link*** function.

**Single predictor logistic regression**

Let the random variable be dichotomous with event Y=1 and event probability . Then, is the failure probability.

A single predictor () logistic regression model is

where

is the log-odds of success,

is called a linear predictor in reference to the predictor

is a logit link function of .

A link function relates an expected response () to a linear function of the predictor ().

**-predictor model**

A -predictor logistic regression model is expressed as

**Interpretation of parameters**

***Continuous predictor***

Given a -predictor logistic regression model , let and be two distinct values of the predictor . Then, the models at and, when all other factors remain fixed, will be

Then,

If , then for every 1 unit increase in the value of , the odds will increase by .

If , then for every *s* units increase in the value of , the odds will increase by , and so on.

**Example 13.7.2**

Let . Then,

(1) for per unit increase in , odd of success will be times the baseline odds of success (**base line odds=1**, always).

(2) 2.5 units increase in , odd of success will be times the baseline odds of success.

***Categorical predictor***

Let the predictor be dichotomous with values for male and for female. The odds of success at and are:

Then,

This implies that the odds of success for females ( is times the odds of success for males ().

If has labels, odds ratios are computed with one label considered as the reference group.

**Notes:**

1. SAS uses MLE method to estimates parameters iteratively.
2. is the odds ratio for a unit increase of the continuous (given all other predictors remain fixed).
3. If is a categorical predictor with labels, odds ratios are computed in SAS with one label as the reference group.

**SAS implementation**

There are two forms of model statement in SAS

1. **events/trials** syntax, with **events** and **trials** separated by a slash (/) when summary information on dichotomous responses on multiple trials are available, e.g.,

**model n/N=x1 x2 x3;**

where n is the number of successes (also called events) out of N trials, and x1 x2 x3 are three predictors separated by space.

1. **single-trial** syntax, with individual information on multiple subjects are available, e.g.,

**model y=x1 x2 x3;**

where y is 1/0 or yes/no response for each individual in the study.

SAS proc logistic fits a logistic regression model. It produces ML estimates of parameters, their standard error and statistics to assess the fit of the model.

**Example 13.7.3: Using a single-trial syntax to predicting task performance**

We are given a dichotomous response with individual level information

and continuous predictor : experience, amount to training in months.

We want to study the effect of computer programming experience on the ability to complete a complex task. We will fit a single-trial syntax logistic regression model:

**data** task;

input experience task @@;

cards;

4 1 5 0 3 0 7 1

4 1 6 0 7 1 3 0

6 1 9 1 8 1 2 0

5 1 1 0 5 1 9 1

3 0 9 1

;

**run**;

\*ods select ResponseProfile GlobalTests ParameterEstimates OddsRatios;

title "logistic reg with desc option";

**proc** **logistic** data=task desc;

model task=experience;

**run**;

By default, proc logistic predicts the log odds of the lowest response value with **reference** to the highest alpha-numeric response value. The **desc** option reverse the default and predicts the log odds for the highest response value (1, here) reference to the lowest response value (0, here).

Alternately, we could specify which value of the response to use as a reference:

\*ods select ResponseProfile GlobalTests ParameterEstimates OddsRatios;

title "logistic reg with ref for response specified";

**proc** **logistic** data=task;

model task (ref="0")=experience;

**run**;

The LOGISTIC Procedure

| **Model Information** | |
| --- | --- |
| **Data Set** | WORK.TASK |
| **Response Variable** | task |
| **Number of Response Levels** | 2 |
| **Model** | binary logit |
| **Optimization Technique** | Fisher's scoring |

|  |  |
| --- | --- |
| **Number of Observations Read** | 18 |
| **Number of Observations Used** | 18 |

| **Response Profile** | | |
| --- | --- | --- |
| **Ordered Value** | **task** | **Total Frequency** |
| **1** | 1 | 11 |
| **2** | 0 | 7 |

|  |
| --- |
| **Probability modeled is task=1.** |

| **Model Convergence Status** |
| --- |
| Convergence criterion (GCONV=1E-8) satisfied. |

| **Model Fit Statistics** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 26.057 | 17.361 |
| **SC** | 26.947 | 19.142 |
| **-2 Log L** | 24.057 | 13.361 |

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 10.6961 | 1 | 0.0011 |
| **Score** | 8.3122 | 1 | 0.0039 |
| **Wald** | 4.5386 | 1 | 0.0331 |

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -4.5915 | 2.3161 | 3.9298 | 0.0474 |
| **experience** | 1 | 1.0509 | 0.4933 | 4.5386 | 0.0331 |

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **experience** | 2.860 | 1.088 | 7.521 |

**Notes:**

The model fit statistics Akaike Information Criterion (**AIC**) and Schwarz Criterion (**SC**) are given by

, where *m* is the number of levels of the dependent variable and *k* is the number of predictors in the model. The model with the smallest is the best. In this problem, , .

Therefore,

, where ’s are the frequency values of the *i*th observation, and and were defined previously. The model with the smallest  is the best.

Three Chi-square tests (**Likelihood Ratio**, **Score** and **Wald**) test that at least one of the predictor’s regression coefficient is not equal to zero in the model.

**Proc Logistic basics**

* PROC LOGISTIC and MODEL statements are required.
* CLASS statement MUST precede MODEL statement along with the list categorical variables with respective *v-options* in parentheses, if desired.
  + class x1 (ref= “No”) x2 x7 (ref= “2”) /param=ref;
* Global *v-options* in CLASS statement specified after a slash (/) applies to all class variables. The individual CLASS variable *v-options* override global *v-options*.

**Example 13.7.4: Using an event-trial syntax to predict risk of uninsured**

We wish to assess the risk of uninsured by marital status. Marital status has the following five groups ():

Married, Widowed, Divorced, Separated and Never married.

For response, the event of interest is being uninsured. The table below defines the number of women with an **event** (i.e., count of uninsured women) out of **total** number of women by each group of the marital status variable.

The model provides odds ratio of being uninsured for four groups () of the marital status variable with one group being considered as a reference group.

We need **class** statement to specify that status is a categorical variable. See the SAS code below:

**data** uninsured;

length status $**10**;

input total event status $;

cards;

101294 13451 Married

3310 703 Widowed

20240 4397 Divorced

4929 1282 Separated

65058 15986 Unmarried

;

**run**;

\*ods select ParameterEstimates OddsRatios;

**proc** **logistic** data= uninsured;

class status (ref='Married')/param=ref;

model event/total=status;

**run**;

The **param=ref** option after the slash requests dummy coding, rather than the default effects coding, for the levels of **status**.

With this option specified, the odds ratios are just the exponentiation of the respective parameter estimates.

|  |
| --- |
| The SAS System |

The LOGISTIC Procedure

| **Model Information** | |
| --- | --- |
| **Data Set** | WORK.UNINSURED |
| **Response Variable (Events)** | event |
| **Response Variable (Trials)** | total |
| **Model** | binary logit |
| **Optimization Technique** | Fisher's scoring |

|  |  |
| --- | --- |
| **Number of Observations Read** | 5 |
| **Number of Observations Used** | 5 |
| **Sum of Frequencies Read ()** | 194831 |
| **Sum of Frequencies Used** | 194831 |

| **Response Profile** | | |
| --- | --- | --- |
| **Ordered Value** | **Binary Outcome** | **Total Frequency** |
| **1** | Event | 35819 |
| **2** | Nonevent | 159012 |

| **Class Level Information** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Class** | **Value** | **Design Variables** | | | |
| **status** | **Divorced** | 1 | 0 | 0 | 0 |
|  | **Married** | 0 | 0 | 0 | 0 |
|  | **Separated** | 0 | 1 | 0 | 0 |
|  | **Unmarried** | 0 | 0 | 1 | 0 |
|  | **Widowed** | 0 | 0 | 0 | 1 |

| **Model Convergence Status** |
| --- |
| Convergence criterion (GCONV=1E-8) satisfied. |

| **Model Fit Statistics** | | | |
| --- | --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** | |
| **Log Likelihood** | **Full Log Likelihood** |
| **AIC** | 185939.47 | 182166.43 | 59.267 |
| **SC** | 185949.65 | 182217.33 | 110.166 |
| **-2 Log L** | 185937.47 | 182156.43 | 49.267 |

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 3781.0426 | 4 | <.0001 |
| **Score** | 3778.9174 | 4 | <.0001 |
| **Wald** | 3685.7557 | 4 | <.0001 |

| **Type 3 Analysis of Effects** | | | |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **status** | 4 | 3685.7557 | <.0001 |

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | -1.8765 | 0.00926 | 41074.6662 | <.0001 |
| **status** | **Divorced** | 1 | 0.5947 | 0.0194 | 939.8945 | <.0001 |
| **status** | **Separated** | 1 | 0.8310 | 0.0338 | 605.7987 | <.0001 |
| **status** | **Unmarried** | 1 | 0.7549 | 0.0130 | 3379.0164 | <.0001 |
| **status** | **Widowed** | 1 | 0.5659 | 0.0435 | 169.2801 | <.0001 |

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **status Divorced vs Married** | 1.812 | 1.745 | 1.883 |
| **status Separated vs Married** | 2.296 | 2.149 | 2.453 |
| **status Unmarried vs Married** | 2.127 | 2.074 | 2.182 |
| **status Widowed vs Married** | 1.761 | 1.617 | 1.918 |

**AIC and BIC computation from given values of**

In this problem, , . Therefore,

Three Chi-square tests (**Likelihood Ratio**, **Score** and **Wald**) test that at least one of the predictor’s regression coefficient is not equal to zero in the model.

LR Chi-Square statistic= -2 Log  L(null model) - 2 Log L(fitted model) , where L(null model) refers to the **Intercept Only** model and L(fitted model) refers to the **Intercept and Covariates** model.

**Example 13.7.5: Assessing risk of child smoking by parental smoking**

Parental smoking status- both smoke, one smokes, neither smokes

Child smoking status: yes (smoke), no (no smoke).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | |  | | --- | |  | | | | **Table of parent by child** | | | | | --- | --- | --- | --- | | **Parent smokes** | **Child smokes** | | | | **yes** | **no** | **Total** | | **both** | |  | | --- | | 400 | | |  | | --- | | 1380 | | |  | | --- | | 1780 | | | **one** | |  | | --- | | 416 | | |  | | --- | | 1823 | | |  | | --- | | 2239 | | | **neither** | |  | | --- | | 188 | | |  | | --- | | 1168 | | |  | | --- | | 1356 | | | **Total** | |  | | --- | | 1004 | | |  | | --- | | 4371 | | |  | | --- | | 5375 | | |

Find the odds-ratio of smoking of child as compared to parents smoking.

**data** smoke;

input parent $ child $ count;

cards;

both yes 400

both no 1380

one yes 416

one no 1823

neither yes 188

neither no 1168

;

**proc** **logistic** data=smoke;

freq count;

class parent (ref="neither")/param=ref;

model child (ref="no")= parent;

**run**;

The LOGISTIC Procedure

| **Model Information** | |
| --- | --- |
| **Data Set** | WORK.SMOKE |
| **Response Variable** | child |
| **Number of Response Levels** | 2 |
| **Frequency Variable** | count |
| **Model** | binary logit |
| **Optimization Technique** | Fisher's scoring |

|  |  |
| --- | --- |
| **Number of Observations Read** | 6 |
| **Number of Observations Used** | 6 |
| **Sum of Frequencies Read** | 5375 |
| **Sum of Frequencies Used** | 5375 |

| **Response Profile** | | |
| --- | --- | --- |
| **Ordered Value** | **child** | **Total Frequency** |
| **1** | no | 4371 |
| **2** | yes | 1004 |

|  |
| --- |
| **Probability modeled is child='yes'.** |

| **Class Level Information** | | | |
| --- | --- | --- | --- |
| **Class** | **Value** | **Design Variables** | |
| **parent** | **both** | 1 | 0 |
|  | **neither** | 0 | 0 |
|  | **one** | 0 | 1 |

| **Model Convergence Status** |
| --- |
| Convergence criterion (GCONV=1E-8) satisfied. |

| **Model Fit Statistics** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 5178.510 | 5144.144 |
| **SC** | 5185.100 | 5163.913 |
| **-2 Log L** | 5176.510 | 5138.144 |

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 38.3658 | 2 | <.0001 |
| **Score** | 37.5663 | 2 | <.0001 |
| **Wald** | 37.0861 | 2 | <.0001 |

| **Type 3 Analysis of Effects** | | | |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **parent** | 2 | 37.0861 | <.0001 |

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | -1.8266 | 0.0786 | 540.2949 | <.0001 |
| **parent** | **both** | 1 | 0.5882 | 0.0970 | 36.8105 | <.0001 |
| **parent** | **one** | 1 | 0.3491 | 0.0955 | 13.3481 | 0.0003 |

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **parent both vs neither** | 1.801 | 1.489 | 2.178 |
| **parent one vs neither** | 1.418 | 1.176 | 1.710 |