Multiple logistic regression

Biometry 755

Spring 2009

Multivariable logistic regression

So far our study of logistic regression has been restricted to models containing a single covariate. We want to extend these methods to allow for multiple regressors.

Suppose we have a collection of k independent variables $X_1, \ldots, X_k$ and binary outcome variable Y. The multiple logistic regression model is

$$\ln\left[\frac{\text{Prob}(Y=1|X_1,\ldots,X_k)}{1-\text{Prob}(Y=1|X_1,\ldots,X_k)}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

IMPACT study example

We will model the log odds of remaining drug free for 12 months as a linear function of age (AGE), the number of prior drug treatments (NDRUGTX), IV drug use history (IVHX), treatment arm (TREAT), and treatment site (SITE).

```
proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site;
run;
quit;
```


Logistic output - model fit

```
         Model Fit Statistics
                                Intercept
                 Intercept            and
Criterion             Only     Covariates

AIC                655.729        634.262
SC                 660.083        664.743
-2 Log L           653.729        620.262

        Testing Global Null Hypothesis: BETA=0
Test                     Chi-Square       DF     Pr > ChiSq

Likelihood Ratio            33.4668        6         <.0001
Score                       31.6135        6         <.0001
Wald                        29.7216        6         <.0001
```

Logistic output - covariate assessment

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	9.2074	0.0024
ndrugtx	1	5.9312	0.0149
ivhx	2	11.1363	0.0038
treat	1	5.2475	0.0220
site	1	0.3266	0.5677

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3726	0.5526	18.4307	<.0001
age		1	0.0522	0.0172	9.2074	0.0024
ndrugtx		1	-0.0624	0.0256	5.9312	0.0149
ivhx	Previous	1	-0.6350	0.2857	4.9402	0.0262
ivhx	Recent	1	-0.7860	0.2471	10.1210	0.0015
treat		1	0.4553	0.1988	5.2475	0.0220
site		1	0.1231	0.2155	0.3266	0.5677

Covariate assessment (cont.)

Just as in multiple linear regression, we must take care in interpretation of significant covariate effects for multiple logistic regression. For example, the covariate NDRUGTX is significant ($p = 0.0149$) so we conclude that the number of previous drug treatments contributes significantly to a model already containing AGE, IVHX, TREAT and SITE. We could also say that the number of previous drug treatments significantly explains the observed variability in the log odds of remaining drug free for 12 months, after adjusting for the effects of AGE, IVHX, TREAT and SITE.

The fitted model

The fitted model is

$$\ln\left[\frac{\text{Prob(DFREE} = 1)}{1 - \text{Prob(DFREE} = 1)}\right] = -2.37 + 0.052 \times \text{AGE}$$
$$-0.062 \times \text{NDRUGTX} - 0.64 \times \text{IVH}$$
$$-0.79 \times \text{IVHX}_2 + 0.46 \times \text{TREAT}$$
$$+0.12 \times \text{SITE}$$

Note that the conditional notation is suppressed for convenience, but is understood to be present.

ORs

Suppose we want to use the fitted model to compute the odds ratio for remaining drug free for 12 months for: 40 year-olds, with no previous drug treatments, recent drug use history, randomized to the short treatment arm, and at site A, relative to 30 year-olds with all the same covariate values.

ORs (cont.)

For the 40-year olds, we have

$$\begin{aligned} \ln(\text{odds DFREE } = 1) &= -2.37 + 0.052 \times 40 - 0.062 \times 0 \\ &\quad -0.64 \times 0 - 0.79 \times 1 + 0.46 \times 0 + 0.12 \times \\ &= -1.08 \end{aligned}$$

For the 30-year olds, we have

$$\begin{aligned} \ln(\text{odds DFREE } = 1) &= -2.37 + 0.052 \times 30 - 0.062 \times 0 \\ &\quad -0.64 \times 0 - 0.79 \times 1 + 0.46 \times 0 + 0.12 \times \\ &= -1.60 \end{aligned}$$

ORs (cont.)

Therefore,

$$\begin{aligned} \ln(\text{odds ratio DFREE } = 1) &= -1.08 -^{-} 1.60 \\ &= 0.52 \end{aligned}$$

so that

$$\text{odds ratio DFREE } = 1 \quad = \quad e^{0.52} = 1.68$$

ORs (cont.)

Now suppose we want to use the fitted model to compute the odds ratio for remaining drug free for 12 months for: 40 year-olds, with no previous drug treatments, recent drug use history, randomized to the long treatment arm, and at site B, relative to 30 year-olds with all the same covariate values.

ORs (cont.)

For the 40-year olds, we have

$$\begin{aligned}\ln(\text{odds DFREE } = 1) &= -2.37 + 0.052 \times 40 - 0.062 \times 0 \\ &\quad -0.64 \times 0 - 0.79 \times 1 + 0.46 \times 1 + 0.12 \times \\ &= -0.5\end{aligned}$$

For the 30-year olds, we have

$$\begin{aligned}\ln(\text{odds DFREE } = 1) &= -2.37 + 0.052 \times 30 - 0.062 \times 0 \\ &\quad -0.64 \times 0 - 0.79 \times 1 + 0.46 \times 1 + 0.12 \times \\ &= -1.02\end{aligned}$$

ORs (cont.)

Therefore,

$$\ln(\text{odds ratio DFREE } = 1) = -0.5 -^{-} 1.02$$
$$= 0.52$$

so that

$$\text{odds ratio DFREE } = 1 = e^{0.52} = 1.68$$

Hmmmmmmm...

Why did the ORs come out the same even though the covariate values were different?

General rule for ORs

The following general rules apply to multiple logistic regression models.

- For a continuous covariate X, $e^{\Delta X \cdot \beta}$ is the OR of $Y = 1$ for subjects with a difference of ΔX in their covariate values, *where all other covariate values are the same*.
- For a categorical covariate, Z, with p levels, and $p - 1$ corresponding indicators $X_1, X_2, \ldots, X_{p-1}$, e^{β_j} $(j = 1, \ldots, p-1)$ is the OR of $Y = 1$ for the group represented by indicator X_j relative to the reference group, *where all other covariate values are the same*.

Default ORs in SAS

By default, SAS will produce all the ORs and corresponding 95% CIs for categorical covariates comparing all levels to the reference category, and for all continuous covariates with a one-unit change in the covariate. If you want an OR and 95% CI for a change in a continuous covariate other than a unit increase, you need to use the UNITS statement in PROC LOGISTIC.

See if you can interpret these ORs correctly.

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.054	1.019	1.090
ndrugtx	0.940	0.894	0.988
ivhx Previous vs Never	0.530	0.303	0.928
ivhx Recent vs Never	0.456	0.281	0.740
treat	1.577	1.068	2.328
site	1.131	0.741	1.725

Interpreting ORs

AGE

NDRUGTX

IVHX: PREVIOUS VS NEVER

IVHX: RECENT VS NEVER

TREAT

SITE

Estimating probabilities

What is the probability that a 30 year-old with 5 prior drug treatments, a history of recent IV drug use, who is randomized to the long treatment arm at site A remains drug free for 12 months?

Recall that

$$\hat{\pi} = \frac{e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}}.$$

Interaction in logistic regression

You can fit a model with interaction just as you did in multiple linear regression. In the current example, it would be worthwhile to investigate potential interaction between the variable TREAT and all other covariates since our primary objective is to understand the relationship between the probability of remaining drug free and the treatment arm (short or long). Failure to account for any significant interaction would result in interpretations that don't accurately depict that relationship.

For example, consider the model with an interaction between AGE and TREAT.

IMPACT example with interaction

```
proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site age*treat;
run;
```

```
                    Odds Ratio Estimates
                             Point       95% Wald
Effect                    Estimate   Confidence Limits
ndrugtx                      0.938   0.892        0.987
ivhx    Previous vs Never    0.526   0.300        0.923
ivhx    Recent   vs Never    0.457   0.281        0.742
site                         1.110   0.727        1.696
```

ORs in the presence of interaction

What happened to the ORs for AGE and TREAT in the default output? Since the model included an interaction term between AGE and TREAT, it is *NOT* appropriate to report a single adjusted OR and CI for each covariate. An interaction term means that the effect of treatment arm on the log odds of remaining drug free depends on the subject's age, and vice versa.

SAS output with interaction

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.4339	0.8103	3.1310	0.0768
age		1	0.0239	0.0250	0.9133	0.3392
ndrugtx		1	-0.0639	0.0257	6.1763	0.0129
ivhx	Previous	1	-0.6424	0.2868	5.0176	0.0251
ivhx	Recent	1	-0.7837	0.2473	10.0420	0.0015
treat		1	-1.1775	1.0600	1.2340	0.2666
site		1	0.1044	0.2162	0.2331	0.6292
age*treat		1	0.0500	0.0320	2.4448	0.1179

Calculating ORs in the presence of interaction

Suppose I want to estimate an OR and 95% CI for the odds of remaining drug free for 12 months comparing those in the long arm to those in the short arm for 40 year-old subjects, assuming other covariates are the same. Keeping in mind that TREAT = 1 for the long arm and TREAT = 0 for the short arm, we have

ORs in the presence of interaction (cont.)

$$\ln\left[\frac{\text{odds (DFREE } = 1|\text{TREAT } = 1, \text{AGE } = 40)}{\text{odds (DFREE } = 1|\text{TREAT } = 0, \text{AGE } = 40)}\right] =$$

Getting SAS to construct it for you ...

```
proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site age*treat;
    contrast '40: long vs. short arm'
    treat 1 age*treat 40/estimate = exp;
run;
```

```
        Contrast Test Results
Contrast                DF  Wald Chi-Square  Pr > ChiSq

40: long vs. short arm   1           7.0293      0.0080
        Contrast Rows Estimation and Testing Results
                                               Standard
Contrast                Type   Row  Estimate   Error
40: long vs. short arm   EXP     1    2.2786  0.7078

Alpha   Confidence Limits
0.05    1.2395      4.1887
```


Confounding

We assess confounding in multiple logistic regression in a manner similar to that used in multiple linear regression. From an analytic viewpoint, if the inclusion of a secondary covariate (Z) in the model *meaningfully* changes the parameter estimate for the exposure covariate (X), then Z is said to confound the relationship between the outcome Y and the exposure X. In practical terms, adjusting for Z meaningfully changes the relationship between Y and X. (NOTE: Remember that you should always assess interaction before confounding.)

Confounding (cont.)

Let's assess whether race confounds the relationship between treatment arm of the trial and the log odds of remaining drug free for 12 months. We begin by assessing whether there is significant interaction between RACE and TREAT.

```
proc logistic data = one descending;
  class ivhx (param = ref ref = 'Never');
  model dfree = age ndrugtx ivhx treat site race race*treat;
run;
quit;
```


Confounding (cont.)

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	8.2955	0.0040
ndrugtx	1	5.4938	0.0191
ivhx	2	9.5195	0.0086
treat	1	5.9676	0.0146
site	1	0.5440	0.4608
race	1	2.2265	0.1357
treat*race	1	1.1631	0.2808

The interaction effect of TREAT and RACE is not significant so we move on to assessing the presence of confounding.

Confounding (cont.)

```
*MODEL WITHOUT RACE PRESENT;
proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site;
run;
quit;

*MODEL WITH RACE PRESENT;
proc logistic data = one descending;
    class ivhx (param = ref ref = 'Never');
    model dfree = age ndrugtx ivhx treat site race
run;
quit;
```


Confounding (cont.)

We look at the variable TREAT to determine if there is a meaningful change in the estimated parameter comparing the fitted models with and without RACE.

Confounding (cont.)

Output for analysis without RACE.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3726	0.5526	18.4307	<.0001
age		1	0.0522	0.0172	9.2074	0.0024
ndrugtx		1	-0.0624	0.0256	5.9312	0.0149
ivhx	Previous	1	-0.6350	0.2857	4.9402	0.0262
ivhx	Recent	1	-0.7860	0.2471	10.1210	0.0015
treat		1	0.4553	0.1988	5.2475	0.0220
site		1	0.1231	0.2155	0.3266	0.5677

Confounding (cont.)

Output for analysis with RACE.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.4054	0.5548	18.7975	<.0001
age		1	0.0504	0.0173	8.4550	0.0036
ndrugtx		1	-0.0615	0.0256	5.7559	0.0164
ivhx	Previous	1	-0.6033	0.2872	4.4118	0.0357
ivhx	Recent	1	-0.7327	0.2523	8.4328	0.0037
treat		1	0.4425	0.1993	4.9302	0.0264
site		1	0.1486	0.2172	0.4681	0.4939
race		1	0.2261	0.2233	1.0251	0.3113

Confounding (cont.)

There is not a meaningful change in the parameter estimate for TREAT comparing the models with and without RACE. Therefore race does not confound the effect of treatment arm on the log odds of remaining drug free for 12 months, while controlling for the effects of the other covariates.