Simple logistic regression

Biometry 755

Spring 2009

Model assumptions

1. The observed data are independent realizations of a binary response variable Y that follows a Bernoulli distribution.
2. The logit of Prob$(Y = 1|X)$ is linear in X.

IMPACT study

Investigate the relationship between a subject’s treatment randomization arm and the subject’s ability to remain drug-free for 12 months.

TREAT Treatment randomization assignment
(0 = Short, 1 = Long)

DFREE Remained drug free for 12 months
(0 = No, 1 = Yes)

SAS code

```
proc freq data = one;
    tables treat*dfree/missing;
run;

proc logistic data = one descending;
    class treat (param = ref ref = ’Long’);
    model dfree = treat;
run;
```

Logistic output

```
The LOGISTIC Procedure

Model Information

Data Set                    WORK.ONE
Response Variable           dfree   Remained drug free
                                    for 12 months
Number of Response Levels   2
Model                       binary logit
Optimization Technique      Fisher's scoring



Number of Observations Read         575
Number of Observations Used         575
```

Logistic output (cont.)

```
               Response Profile
 Ordered                                  Total
   Value     dfree                    Frequency


       1     Remained drug free             147
       2     Otherwise                      428

Probability modeled is dfree='Remained drug free'.

   Class Level Information
                        Design
Class     Value     Variables
treat     Long              1
          Short             0
```

Logistic output (cont.)

```
                    Model Convergence Status


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


         Model Fit Statistics
                             Intercept
              Intercept            and
Criterion          Only     Covariates


AIC             655.729        652.551
SC              660.083        661.259
-2 Log L        653.729        648.551
```

Logistic output (cont.)

```
        Testing Global Null Hypothesis: BETA=0


Test                   Chi-Square       DF      Pr > ChiSq


Likelihood Ratio           5.1782        1          0.0229
Score                      5.1626        1          0.0231
Wald                       5.1266        1          0.0236
```

Logistic output (cont.)

```
        Type 3 Analysis of Effects
                        Wald
Effect      DF    Chi-Square    Pr > ChiSq

treat        1        5.1266        0.0236

                Analysis of Maximum Likelihood Estimates
                                     Standard          Wald
Parameter          DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept           1     -1.2978      0.1433       82.0211        <.0001
treat     Long      1      0.4371      0.1931        5.1266        0.0236
```

Logistic output (cont.)

```
              Odds Ratio Estimates

                          Point       95% Wald
Effect                 Estimate   Confidence Limits
treat Long  vs Short      1.548   1.060       2.260
```

Assessing significance of covariates

For each covariate in the model, the output provides an estimated parameter, the standard error of that estimate, a test-statistic, and a p-value. One of the most common tests used to assess the significance of MLEs is the *Wald test*. The test associated with a single parameter is based on the following null hypothesis:

$$\begin{aligned} H_0 : \beta_j &= 0 \ (\text{all other } \beta\text{s are non-zero}) \\ H_A : \beta_j &\neq 0 \ (\text{all other } \beta\text{s are non-zero}) \end{aligned}$$

Assessing significance of covariates (cont.)

The Wald test is obtained by comparing the MLE of the slope parameter, $\hat{\beta}_j$, to an estimate of its standard error. The resulting ratio, under the hypothesis that $\beta_j = 0$, follows a standard normal distribution. That is

$$\frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim \text{Normal}(0, 1).$$

Recall that the square of a Normal(0,1) random variable has a chi-square distribution with one degree of freedom. The SAS output (shown on Slide 9) provides the chi-square version of the Wald test along with corresponding p-values. We conclude that treatment arm is significantly associated with the probability of being drug free for 12 months ($p = 0.0236$).

Interpreting model parameters

Using the MLEs shown on Slide 9, the fitted model is

$$\begin{aligned}\ln\left[\frac{\widehat{\text{Prob}}(\text{DFREE}=1|\text{TREAT})}{1-\widehat{\text{Prob}}(\text{DFREE}=1|\text{TREAT})}\right] &= \hat{\beta}_0+\hat{\beta}_1\times\text{TREAT}\\ &= -1.2978+0.4371\times\text{TREAT}\end{aligned}$$

How do we interpret these parameter estimates?

Interpreting model parameters (cont.)

$$\ln\left[\frac{\widehat{\text{Prob}}(\text{DFREE}=1|\text{TREAT})}{1-\widehat{\text{Prob}}(\text{DFREE}=1|\text{TREAT})}\right] = -1.2978+0.4371\times\text{TREAT}$$

- TREAT $=0$ (Short arm) $\Rightarrow$
  $\ln\left[\frac{\widehat{\text{Prob}}(\text{DFREE = 1}|\text{TREAT=0})}{1-\widehat{\text{Prob}}(\text{DFREE = 1}|\text{TREAT=0})}\right]=-1.2978.$

In words, the log odds of remaining drug free for 12 months among subjects randomized to the short arm of the trial is -1.2978. Therefore, the odds of remaining drug free for 12 months among subjects randomized to the short arm of the trial is $e^{-1.2978}\doteq 0.27$.

Interpreting model parameters (cont.)

$$\ln\left[\frac{\widehat{\text{Prob}}(\text{DFREE} = 1|\text{TREAT})}{1-\widehat{\text{Prob}}(\text{DFREE} = 1|\text{TREAT})}\right] \quad = \quad -1.2978 + 0.4371 \times \text{TREAT}$$

- TREAT $= 1$ $\quad$ (Long arm) $\Rightarrow$
  $$\ln\left[\frac{\widehat{\text{Prob}}(\text{DFREE = 1}|\text{TREAT=1})}{1-\widehat{\text{Prob}}(\text{DFREE = 1}|\text{TREAT=1})}\right] = -1.2978 + 0.4371.$$

The log odds of remaining drug free for 12 months among subjects randomized to the long arm of the trial is -1.2978 + 0.4371 = -0.8606. Therefore, the odds of remaining drug free for 12 months among subjects randomized to the long arm of the trial is $e^{-0.8606} \doteq 0.42$.

Comparing groups

Usually we wish to compare the two groups using an odds ratio. Since we have the odds of remaining drug free for 12 months for each group, then it is a simple matter to construct the odds ratio.

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

$$\ln(\text{odds (DFREE = 1}|\text{TREAT = 1}) - \ln(\text{odds (DFREE = 1}|\text{TREAT = 0}) =$$

$$-0.8606 - (-1.2978) = 0.4372$$

Then $e^{0.4372} \doteq 1.55 \Rightarrow$ there is a 55% increase in the odds of remaining drug free for 12 months for subjects in the long versus the short arm of the trial.

Useful laws of logarithms

To construct the odds ratio comparing subjects in the long arm to those in the short arm, we relied upon a fundamental property of logarithms. Namely

$$\log\left(\frac{a}{b}\right) = \log a - \log b.$$

This property will be essential to our ability to manipulate estimated βs to construct meaningful odds ratios.

Simple logistic with binary covariate

For the logistic model with binary covariate X,

$$\ln\left[\frac{\text{Prob}(Y=1|X)}{1-\text{Prob}(Y=1|X)}\right] = \beta_0 + \beta_1 X$$

- $\beta_0 + \beta_1 =$ <u>log odds</u> of the event when $X = 1$.
- $\beta_0 =$ <u>log odds</u> of the event when $X = 0$.
- Then the log odds ratio of the event when $X = 1$ relative to $X = 0$ is the difference in the log odds (Slide 17).
- Since $(\beta_0 + \beta_1) - \beta_0 = \beta_1$, then $\beta_1 =$ <u>log odds</u> ratio of the event when $X = 1$ relative to when $X = 0$.
- $e^{\beta_1} =$ <u>odds ratio</u> of the event when $X = 1$ relative to when $X = 0$.

Interpreting odds ratios

Suppose an odds ratio compares group A (numerator) to group B (denominator). Then:

- An odds ratio between 0 and 1 is interpreted as a percent reduction in the odds of the event. For example, an odds ratio of 0.7 is interpreted as a 30% reduction in the odds of the event for those in group A relative to those in group B.
- An odds ratio between 1 and 2 is interpreted as a percent increase in the odds of the event. For example, an odds ratio of 1.6 is interpreted as a 60% increase in the odds of the event for those in group A relative to those in group B.

Interpreting odds ratios (cont.)

- An odds ratio greater than 2 is interpreted as a multiplicative increase in the odds of the event. For example, an odds ratio of 2.1 is interpreted as follows: “The odds of the event for subjects in group A are approximately twice that of the subjects in group B.”

OR confidence intervals

Recall that the general form for a 95% confidence interval is

$$\hat{\theta} \pm 1.96 \times \widehat{\text{SE}}(\hat{\theta}).$$

The formula for the confidence interval of the odds ratio follows this same structure. For the case where $\widehat{\text{OR}} = e^{\hat{\beta}_j}$, an approximate 95% CI for the OR is

$$e^{\{\hat{\beta}_j \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_j)\}}.$$

OR confidence intervals (cont.)

On Slide 9, we saw that $\hat{\beta}_{\text{TREAT}} \doteq 0.4371$ and $\widehat{\text{SE}}(\hat{\beta}_{\text{TREAT}}) \doteq 0.1931$. Therefore, a 95% confidence interval for the odds ratio of DFREE for subjects in the long arm relative to those in the short arm (estimated on Slide 16), is

$$e^{\{0.4371 \pm 1.96 \times 0.1931\}} = (e^{0.058624}, e^{0.815576}) \doteq (1.06, 2.26).$$

Fortunately, SAS provides estimated ORs and corresponding CIs by default for ORs of this form (i.e. those that are equivalent to $e^{\hat{\beta}_j}$). Compare our hand-calculated results with the output presented on Slide 10.

IMPACT example with polytomous covariate

We now wish to investigate the relationship between a subject’s IV drug use history and the subject’s ability to remain drug-free for 12 months.

SAS code

```
proc freq data = one;
    tables ivhx*dfree/missing;
run;

proc logistic data = one descending;
    class ivhx (param = ref ref = ’Never’);
    model dfree = ivhx;
run;
```

Logistic output

```
    Class Level Information

                             Design
Class      Value           Variables

ivhx       Never             0       0
           Previous          1       0
           Recent            0       1
```

Logistic output (cont.)

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


AIC             655.729        646.376
SC              660.083        659.440
-2 Log L        653.729        640.376

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


Likelihood Ratio           13.3525        2         0.0013
Score                      13.4161        2         0.0012
Wald                       13.1585        2         0.0014
```

Logistic output (cont.)

```
        Type 3 Analysis of Effects
                        Wald
Effect       DF    Chi-Square    Pr > ChiSq

ivhx          2       13.1585        0.0014

    Analysis of Maximum Likelihood Estimates
                             Standard        Wald
Parameter        DF  Estimate    Error  Chi-Square Pr > ChiSq

Intercept         1   -0.6797   0.1417   22.9977   <.0001
ivhx   Previous  1   -0.4810   0.2657    3.2773   0.0702
ivhx   Recent    1   -0.7748   0.2166   12.7997   0.0003
```

Type 3 analyses

The *Type 3 Analysis* in the PROC LOGISTIC output (Slide 27) is akin to a multiple partial F-test from PROC REG. To review, a Type 3 analysis assesses the significance of the (complete) categorical (ordinal or nominal) variable, while adjusting for the presence of all other variables in the model. For class variables, you should always assess the overall significance of the covariate using a Type 3 analysis. The Wald tests presented under the heading of *Analysis of Maximum Likelihood Estimates* only assess the significance of the dummy variables constructed to fit the model.

The fitted model

The fitted model is

$$\ln\left[\frac{\widehat{\text{Prob}}(\text{DFREE = 1}|\text{IVHX})}{1-\widehat{\text{Prob}}(\text{DFREE = 1}|\text{IVHX})}\right] =$$

$$-0.6797 - 0.4810 \cdot \text{IVHX (previous)} - 0.7748 \cdot \text{IVHX (recent)}$$

Interpreting model parameters

- The log odds of remaining drug free for 12 months for those with no history of IV drug use is -0.6797.
- The log odds of remaining drug free for 12 months for those with a previous history of IV drug use is -0.6797 - 0.4810 = -1.1607.
- The log odds of remaining drug free for 12 months for those with a recent history of IV drug use is -0.6797 - 0.7748 = -1.4545.

Interpreting model parameters (cont.)

Using the property of logarithms from Slide 17, we can easily construct odds ratios.

- The log odds ratio of remaining drug free for 12 months for those with a previous history of IV drug use relative to those with no history of IV drug use is (-0.6797 - 0.4810) - (-0.6797) = -0.4810 $\Rightarrow$ the odds ratio of remaining drug free for 12 months for those with a previous history of IV drug use relative to those with no history of IV drug is $e^{-0.4810} \doteq 0.62$.
- There is a 38% reduction in the odds of remaining drug free for 12 months for those with a previous history of IV drug use relative to those with no history of IV drug.

Interpreting model parameters (cont.)

- The log odds ratio of remaining drug free for 12 months for those with a recent history of IV drug use relative to those with no history of IV drug use is (-0.6797 - 0.7748) - (-0.6797) = -0.7748 $\Rightarrow$ the odds ratio of remaining drug free for 12 months for those with a recent history of IV drug use relative to those with no history of IV drug is $e^{-0.7748} \doteq 0.46$.
- There is a 54% reduction in the odds of remaining drug free for 12 months for those with a recent history of IV drug use relative to those with no history of IV drug.

Default odds ratios and CIs in SAS

The default output for SAS is to provide only those ORs and corresponding CIs that compare the groups represented by the dummy variables to the reference group.

```
        Odds Ratio Estimates

                         Point        95% Wald
Effect                   Estimate  Confidence Limits

ivhx Previous vs Never   0.618     0.367        1.041
ivhx Recent   vs Never   0.461     0.301        0.704
```

Summary of model parameters

For the logistic model

$$\ln\left[\frac{\text{Prob}(Y=1|X)}{1-\text{Prob}(Y=1|X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where X_1 and X_2 are dummy variables coded 0/1 representing groups one and two, respectively, for a three-level categorical covariate, Z,

- β_1 is the log odds ratio of the event for group 1 relative to the reference group $\Rightarrow e^{\beta_1}$ is the odds ratio of the event for group 1 relative to the reference group.
- β_2 is the log odds ratio of the event for group 2 relative to the reference group $\Rightarrow e^{\beta_2}$ is the odds ratio of the event for group 2 relative to the reference group.

Using a different reference group

To obtain ORs and corresponding 95% CI in SAS using a different reference group, it is easiest to simply re-run PROC LOGISTIC and specify a different reference group. The code below identifies recent IV drug users as the reference group.

```
proc logistic data = one descending;
    class ivhx (param = ref ref = ’Recent’);
    model dfree = ivhx;
run;
```

Logistic output

```
                      Odds Ratio Estimates

                                Point        95% Wald
Effect                       Estimate    Confidence Limits

ivhx Never    vs Recent         2.170    1.420       3.318
ivhx Previous vs Recent         1.342    0.778       2.314
```

See if you can interpret these ORs.

IMPACT example with continuous covariate

We now wish to investigate the relationship between a subject’s number of prior drug treatments and the subject’s ability to remain drug-free for 12 months.

```
proc logistic data = one descending;
    model dfree = ndrugtx;
run;
```

Logistic output

```
            Analysis of Maximum Likelihood Estimates
                               Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -0.7678      0.1303       34.7133        <.0001
ndrugtx       1     -0.0749      0.0247        9.2203        0.0024

           Odds Ratio Estimates
              Point          95% Wald
Effect     Estimate      Confidence Limits

ndrugtx       0.928       0.884       0.974
```

Interpreting model parameters

The fitted model is

$$\ln\left[\frac{\text{Prob}(\text{DFREE = 1}|\text{NDRUGTX})}{1-\text{Prob}(\text{DFREE = 1}|\text{NDRUGTX})}\right] = -0.7678 - 0.0749 \cdot \text{NDRUGTX}.$$

Suppose you want to compare the odds of remaining drug free for 12 months for those with 10 prior treatments to those with 5 prior treatments.

Interpreting model parameters (cont.)

- The log odds of remaining drug free for 12 months among subjects with 10 prior drug treatments is $-0.7678 - 0.0749 \cdot 10 = -1.5168$.
- The log odds of remaining drug free for 12 months among subjects with 5 prior drug treatments is $-0.7678 - 0.0749 \cdot 5 = -1.1423$.
- The log odds ratio of remaining drug free for 12 months for subjects with 10 versus 5 prior drug treatments is $-1.5168 -^{-} 1.1423 = -0.3745 \Rightarrow$ The odds of remaining drug free for 12 months comparing those with 10 versus 5 prior drug treatments is $e^{-0.3745} \doteq 0.69$. Therefore, there is a 31% reduction in the odds of remaining drug free for 12 months comparing subjects with 10 prior drug treatments to those with 5 prior drug treatments.

Interpreting model parameters (cont.)

The fitted model is

$$\ln\left[\frac{\text{Prob}(\text{DFREE = 1}|\text{NDRUGTX})}{1-\text{Prob}(\text{DFREE = 1}|\text{NDRUGTX})}\right] = -0.7678 - 0.0749 \cdot \text{NDRUGTX}.$$

Suppose you want to compare the odds of remaining drug free for 12 months for those with 25 prior treatments to those with 20 prior treatments.

Interpreting model parameters (cont.)

- The log odds of remaining drug free for 12 months among subjects with 25 prior drug treatments is $-0.7678 - 0.0749 \cdot 25 = -2.6403$.
- The log odds of remaining drug free for 12 months among subjects with 20 prior drug treatments is $-0.7678 - 0.0749 \cdot 20 = -2.2658$.
- The log odds ratio of remaining drug free for 12 months for subjects with 25 versus 20 prior drug treatments is $-2.6403 -^{-} 2.2658 = -0.3745 \Rightarrow$ The odds of remaining drug free for 12 months comparing those with 25 versus 20 prior drug treatments is $e^{-0.3745} \doteq 0.69$. Therefore, there is a 31% reduction in the odds of remaining drug free for 12 months comparing subjects with 25 prior drug treatments to those with 20 prior drug treatments.

That’s strange ...

Why did those odds ratios come out the same?

ORs when the covariate is continuous

For the logistic model

$$\ln\left[\frac{\text{Prob}(Y=1|X)}{1-\text{Prob}(Y=1|X)}\right]=\beta_0+\beta_1 X$$

where X is a continuous covariate, let x_1 and x_2 be two specific values of the covariate X. (For example, if the covariate X is NDRUGTX, then x_1 might be 25 and x_2 might be 20.) Then ...

ORs when the covariate is continuous (cont.)

- $\beta_0 + \beta_1 x_1$ is the log odds of $Y = 1$ for all subjects with covariate value x_1 so that the corresponding odds are $e^{\beta_0+\beta_1 x_1}$
- $\beta_0 + \beta_1 x_2$ are the log odds of $Y = 1$ for all subjects with covariate value x_2 so that the corresponding odds are $e^{\beta_0+\beta_1 x_2}$.
- $(\beta_0 + \beta_1 x_2) - (\beta_0 + \beta_1 x_1) = (x_2 - x_1)\beta_1$ is the log odds ratio of $Y = 1$ for those with covariate value x_2 relative to those with covariate value x_1, so that the corresponding odds ratio is $e^{(x_2-x_1)\beta_1}$.

So why were the ORs the same?

Based on the third bullet item from Slide 45, the OR depends on the *difference* between the two values of the covariate, not on their actual values. Since 10 and 5 prior drug treatments, and 25 and 20 prior drug treatments are both 5 units apart, the estimated OR comparing both pairs of subjects will be

$$e^{5\cdot\hat{\beta}_1} = e^{5\cdot -0.0749} = e^{-0.3745} \doteq 0.69.$$

ORs and CIs in SAS for continuous covariate

```
proc logistic data = one descending;
    model dfree = ndrugtx/clodds=wald;
    units ndrugtx = 5;
run;
```

```
    Wald Confidence Interval for Adjusted Odds Ratios
Effect           Unit   Estimate   95% Confidence Limits
ndrugtx        5.0000      0.688       0.540         0.876
```