Inference for Logistic Regression

EPI 204

Quantitative Epidemiology III

Statistical Models

Evans County, GA Dataset (1963)

- Data are in evans.dat (text, no header), evans.sas7bdat (SAS version 9 dataset), evans.sav (SPSS dataset), and evans.dta (Stata dataset) on the website given in the syllabus.
- The data are from a cohort study in which 609 white males were followed for 7 years, with coronary heart disease as the outcome of interest.
- The variables are given on the next slide

Variable	Description
ID	Subject ID, one observation per subject
CHD	Coronary heart disease (1) or not (0)
CAT	High catecholamine level (1) or not (0)
AGE	Age in years
CHL	Cholesterol level
SMK	Ever smoked (1) or never smoked (0)
ECG	ECG abnormality (1) or not (0)
DBP	Diastolic blood pressure
SBP	Systolic blood pressure
HPT	= 1 if DBP ≥90 or SBP ≥ 160, otherwise = 0
CH	CAT*HPT
CC	CAT*CHL

```
> vars <- c("ID","CHD","CAT","AGE","CHL","SMK","ECG","DBP","SBP","HPT","CH","CC")
> evans <- read.table("evans.dat",header=F,col.names=vars)
> summary(evans)
       ID              CHD              CAT              AGE              CHL        
 Min.   :   21   Min.   :0.0000   Min.   :0.0000   Min.   :40.00   Min.   : 94.0  
 1st Qu.: 4242   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:46.00   1st Qu.:184.0  
 Median : 9751   Median :0.0000   Median :0.0000   Median :52.00   Median :209.0  
 Mean   : 9213   Mean   :0.1166   Mean   :0.2003   Mean   :53.71   Mean   :211.7  
 3rd Qu.:13941   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:60.00   3rd Qu.:234.0  
 Max.   :19161   Max.   :1.0000   Max.   :1.0000   Max.   :76.00   Max.   :357.0  
      SMK              ECG              DBP              SBP        
 Min.   :0.0000   Min.   :0.0000   Min.   : 60.00   Min.   : 92.0  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 80.00   1st Qu.:125.0  
 Median :1.0000   Median :0.0000   Median : 90.00   Median :140.0  
 Mean   :0.6355   Mean   :0.2726   Mean   : 91.18   Mean   :145.5  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:100.00   3rd Qu.:160.0  
 Max.   :1.0000   Max.   :1.0000   Max.   :170.00   Max.   :300.0  
```

```
 HPT               CH                CC
 Min.   :0.0000   Min.   :0.0000   Min.   :  0.00
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  0.00
 Median :0.0000   Median :0.0000   Median :  0.00
 Mean   :0.4187   Mean   :0.1609   Mean   : 39.96
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:  0.00
 Max.   :1.0000   Max.   :1.0000   Max.   :331.00
```

(one of many possible exploratory plots)

```
> plot(SBP ~ AGE, data=evans)
> abline(coef(lm(SBP~AGE,data=evans)),lwd=2)
> abline(h=140,col="blue",lwd=2)
> abline(h=160,col="red",lwd=2)
> title("Systolic Blood Pressure by Age")
```

Wald tests

- Each coefficient has an estimated variance and therefore standard error.
- In general, they coefficients are correlated.
- Confidence intervals and tests for single coefficients are given in the summary(glm()) output or are easily derived.
- We need to do more work to find confidence intervals and tests for differences of coefficients as when a factor has more than two levels.
- The same trick allows us to get confidence intervals and tests for interaction terms.

```
> coef(summary(hyp.glm))
               Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.37766146  0.3801845 -6.2539671 4.001553e-10
smokingYes  -0.06777489  0.2781242 -0.2436857 8.074742e-01
obesityYes   0.69530960  0.2850851  2.4389544 1.472983e-02
snoringYes   0.87193932  0.3975736  2.1931517 2.829645e-02

CI for obesity log odds ratio 0.6953 ± (1.960)(0.2851)
0.6963 ± 0.5588
(0.1365, 1.2541)
CI for odds ratio use exp()
(1.15, 3.50)
> confint.default(glm(hyp.tbl~smoking+obesity+snoring,
                  binomial,hyp))
                  2.5 %     97.5 %
(Intercept) -3.12280942 -1.6325135
smokingYes  -0.61288823  0.4773385
obesityYes   0.13655304  1.2540662
snoringYes   0.09270929  1.6511693
```

```
> juul1.glm <-
glm(menarche~age+tanner,binomial,data=juul1)
> summary(juul1.glm)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.7758     2.7630  -4.986 6.17e-07 ***
age           0.8603     0.2311   3.723 0.000197 ***
tanner2      -0.5211     1.4846  -0.351 0.725609
tanner3       0.8264     1.2377   0.668 0.504313
tanner4       2.5645     1.2172   2.107 0.035132 *
tanner5       5.1897     1.4140   3.670 0.000242 ***
```

The estimated log odds between tanner 5 and tanner 4 is 5.1897 – 2.5645 = 2.6252

How do we make a confidence interval or hypothesis test?

If X and Y are random variables then

$$V(X-Y)=V(X)+V(Y)-2\text{Cov}(X,Y)$$

more generally, if a and b are any numbers

$$V(aX+bY)=a^2V(X)+b^2V(Y)+2ab\text{Cov}(X,Y)$$

If B is a vector of coefficients $B=(\beta_0,\beta_1,\ldots,\beta_p)$

and V is the $(p+1)$ by $(p+1)$ covariance matrix of B,

and b is a vector of length $(p+1)$ of numbers, then

$b^\top B=\sum_{i=0}^{p} b_i\beta_i$ has variance

$$\text{b}^\top Vb=\sum_{i=0}^{p}\sum_{j=0}^{p} b_i b_j v_{ij}$$

```
> c1 <- coef(juul1.glm)
> v1 <- vcov(juul1.glm)
> c1
(Intercept)         age     tanner2     tanner3     tanner4     tanner5
-13.7758129   0.8603095  -0.5210667   0.8264390   2.5645049   5.1896586
> v1
             (Intercept)         age     tanner2    tanner3    tanner4    tanner5
(Intercept)  7.63420854 -0.59398971 -0.08627975  0.2184034  0.4414426  0.6973030
age         -0.59398971  0.05338584 -0.08439333 -0.1117773 -0.1318233 -0.1548192
tanner2     -0.08627975 -0.08439333  2.20415172  1.2019693  1.2336584  1.2700108
tanner3      0.21840340 -0.11177725  1.20196929  1.5319140  1.3012763  1.3494243
tanner4      0.44144258 -0.13182328  1.23365842  1.3012763  1.4816528  1.4075578
tanner5      0.69730303 -0.15481918  1.27001076  1.3494243  1.4075578  1.9993267
> b1 <- c(0,0,0,0,-1,1)
> t(b1) %*% c1
         [,1]
[1,] 2.625154
> t(b1)%*% v1 %*% b1
          [,1]
[1,] 0.6658638
```

$2.625154 \pm (1.960)\sqrt{0.6658638}$

2.625 ± 1.599 or (1.025, 4.224) log odds ratio or (2.787, 68.33) odds ratio

Alternatives

- We can re-run the analysis with a different default level for the categorical variable so that the desired comparison is a single coefficient.

```
relevel {stats}   Reorder Levels of Factor

The levels of a factor are re-ordered so that the
level specified by ref is first and the others are
moved down.

relevel(x, ref, ...)

x     an unordered factor.
ref   the reference level, typically a string.
```

Contrasts

- A contrast is a weighted combination of factor levels in which the weights add up to 1.
- We can change the coding of a factor with five levels from an intercept and four 0/1 variables to a set of contrasts.
- This can be complex and hard to implement.
- There are other R packages that can handle these separately such as `multcomp`.

Interaction Terms

- An interaction means that the effect of one variable depends on the level of another.
- To get a numerical measure of effect such as odds ratio, we need to specify the level of the other one.
- We can set the other variable at the modal level, the median level, or the mean level, or at a variety of levels
- We will illustrate this with the evans data.

```
> summary(glm(CHD~CAT+CHL+HPT+CAT*HPT,binomial,evans))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.575185   0.758746  -6.030 1.64e-09 ***
CAT          2.190774   0.498842   4.392 1.12e-05 ***
CHL          0.008495   0.003256   2.609  0.00909 **
HPT          0.912656   0.324423   2.813  0.00491 **
CAT:HPT     -1.681469   0.600829  -2.799  0.00513 **

Effect of CAT when HPT = 0 is 2.190774 with inferences given on the line.
Effect of CAT when HPT = 1 is 2.191 - 1.681 = 0.5093
We can analyze this using b1 = c(0,1,0,0,1)

> b1 <- c(0,1,0,0,1)
> t(b1) %*% c1
          [,1]
[1,] 0.5093054
> v <- vcov(evans1.glm)
> t(b1) %*% v %*% b1
          [,1]
[1,] 0.1245581
> sqrt(t(b1) %*% v %*% b1)
          [,1]
[1,] 0.3529279
```

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.575185   0.758746  -6.030 1.64e-09 ***
CAT          2.190774   0.498842   4.392 1.12e-05 ***
CHL          0.008495   0.003256   2.609  0.00909 **
HPT          0.912656   0.324423   2.813  0.00491 **
CAT:HPT     -1.681469   0.600829  -2.799  0.00513 **

Effect of CAT when HPT = 0 is 2.190774 with inferences given on the line.
Effect of CAT when HPT = 1 is 2.191 - 1.681 = 0.5093
> b1 <- c(0,1,0,0,1)
> t(b1) %*% c1
[1,] 0.5093054
> v <- vcov(evans1.glm)
> sqrt(t(b1) %*% v %*% b1)
[1,] 0.3529279

Log odds ratio = 0.5093054, odds ratio = 1.664
Log odds ratio CI is 0.5093 ± (1.960)(0.3529)  0.5093 ± 0.6917
(-.1824, 1.2010) and odds ratio CI is (0.83, 3.32)

Catecholamine level has a significant effect on non-hypertensives, but
a smaller and non-significant effect on hypertensives. 41.9% of subjects
are hypertensive, so we could use b1 <- c(0,1,0,0,0.419) for average effect.
```

Homework: Due 4/25/17

- For the Byssinosis data, there are three levels of workspace. Test the hypothesis that each differs from the other (two are in the summary table and the third requires the covariance matrix). Find 95% confidence intervals for the three odds ratios.
- Do the same for the three levels of employment time.
- Fit a model omitting race and sex but including the smoking by workspace interaction. Find the estimated effect of smoking for each of the three workspaces separately, test the hypothesis of no effect, and find 95% confidence intervals for the odds ratios.