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

Input for Evans.dat (tab delimited)

```
vars <- c("ID","CHD","CAT","AGE","CHL","SMK",
          "ECG","DBP","SBP","HPT","CH","CC")
evans <- read.table("evans.dat",header=F,
                    col.names=vars)
```

read.table, read.csv, etc. are all variants that can handle text file input with different defaults.

By default, reads strings as factors, unless stringsAsFactors=F. Often this option is a good idea.

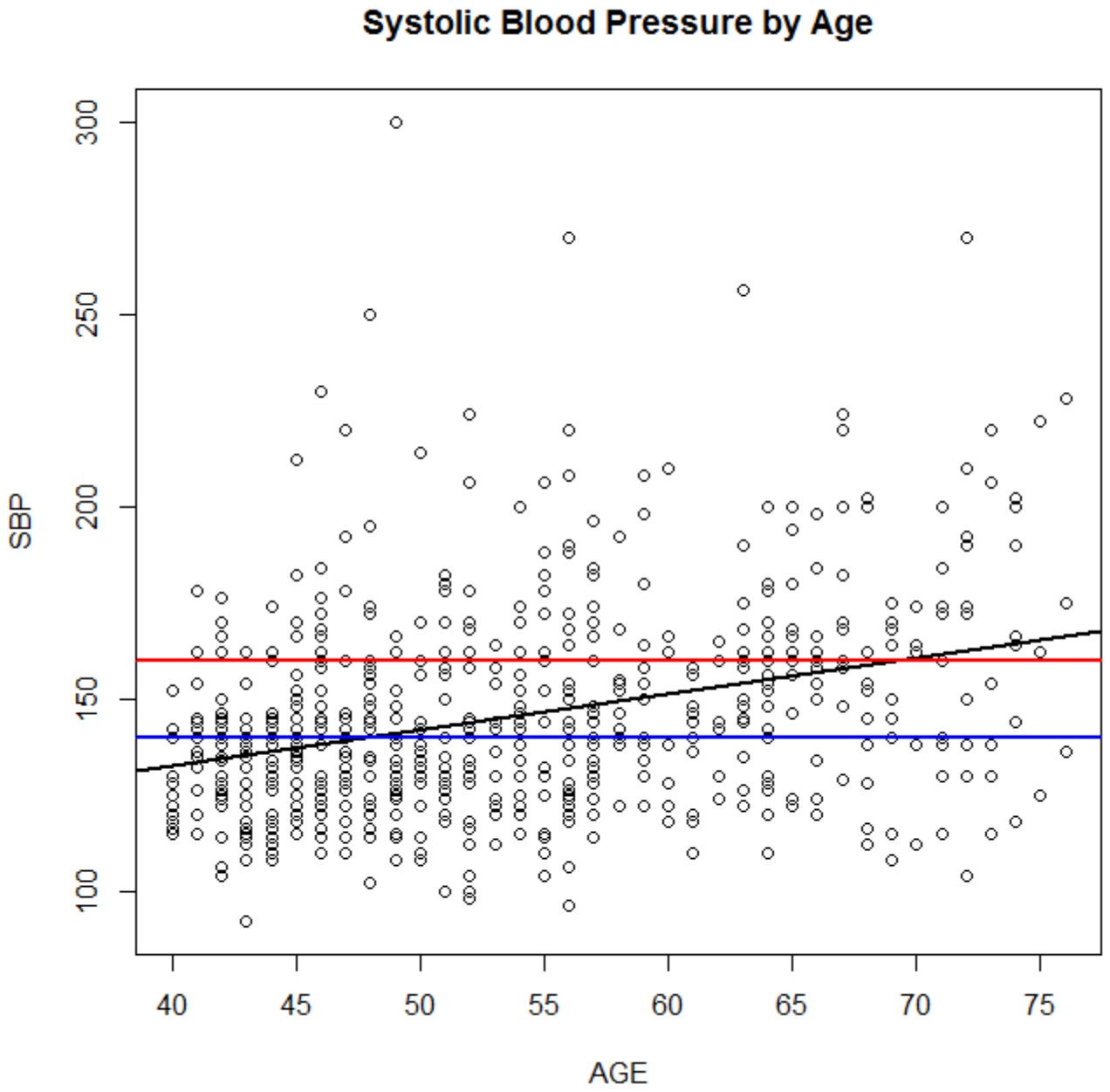
Variable	Description
ID	Subject ID, one observation per subject
CHD	Coronary heart disease (1) or not (0)
CAT	High catecholamine level (1) or not (0)
AGE	Age in years
CHL	Cholesterol level (mg/dL)
SMK	Ever smoked (1) or never smoked (0)
ECG	ECG abnormality (1) or not (0)
DBP	Diastolic blood pressure (mm)
SBP	Systolic blood pressure (mm)
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)  #Systolic Blood Pressure vs. Age
> abline(coef(lm(SBP~AGE,data=evans)),lwd=2)  #Puts regression line on the plot
> abline(h=140,col="blue",lwd=2)  #line at current hypertension threshold
> abline(h=160,col="red",lwd=2)   #line at hypertension threshold used in study
> title("Systolic Blood Pressure by Age")
```


Systolic Blood Pressure by Age
SBP
100
150
200
250
300
40
45
50
55
60
65
70
75
AGE

```
> summary(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.680112   1.136363  -5.879 4.14e-09 ***
CAT          0.715810   0.340180   2.104  0.03536 *
AGE          0.032770   0.015197   2.156  0.03105 *
CHL          0.008608   0.003259   2.641  0.00827 **
SMK          0.802906   0.303001   2.650  0.00805 **
HPT          0.476272   0.289296   1.646  0.09970 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 401.95  on 603  degrees of freedom
AIC: 413.95

Number of Fisher Scoring iterations: 5
```

```
> drop1(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans),test="Chisq")
Single term deletions

Model:
CHD ~ CAT + AGE + CHL + SMK + HPT
       Df Deviance    AIC    LRT Pr(>Chi)
<none>      401.95 413.95
CAT     1   406.33 416.33 4.3805 0.036353 *
AGE     1   406.52 416.52 4.5682 0.032571 *
CHL     1   408.86 418.86 6.9088 0.008577 **
SMK     1   409.65 419.65 7.6990 0.005525 **
HPT     1   404.66 414.66 2.7097 0.099741 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HPT is not statistically significant, but omitting it causes a rise in the AIC, so some might keep it in the model.

Likelihood Ratio Test

- This is used to compare two statistical models that are *nested*, meaning that one (the *full model*) has all the terms of the other (the *reduced model*) plus one or more additional ones.
- For example, the full model might have CHD~CAT+AGE+CHL+SMK+HPT
- And the reduced model might have CHD~CAT+AGE+CHL+SMK (removing HPT) or CHD~CAT+CHL+SMK (removing AGE and HPT)

Likelihood Ratio Test

- If the full model has likelihood L_F and the reduced model has likelihood L_R, then statistical theory says that $-2\ln(L_R/L_F) = -2[\ln(L_R) - \ln(L_F)]$ has approximately a chi-squared distribution with df = the number of omitted variables (with categorical variables counting as one less than the number of categories).
- Since $D = -2(LL - LL_o)$, where LL_o is the log likelihood of the maximal model, we can equally use $D_R - D_F$.
- $D_R - D_F = -2(LL_R - LL_o) - [-2(LL_F - LL_o)]$
  $= -2[\ln(L_R) - \ln(L_F)]$

```
> deviance(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans))
[1] 401.947
> deviance(glm(CHD~CAT+AGE+CHL+SMK,family=binomial,data=evans))
[1] 404.6566
> deviance(glm(CHD~CAT+CHL+SMK,family=binomial,data=evans))
[1] 409.3424
```

Test for omitting HPT from full model uses 404.6566 – 401.947 = 2.7096.
Compare to chi-squared on 1df

```
> 1-pchisq(2.7096,1)
[1] 0.099746
```

(same as produced with drop1)

```
> deviance(glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans))
[1] 401.947
> deviance(glm(CHD~CAT+AGE+CHL+SMK,family=binomial,data=evans))
[1] 404.6566
> deviance(glm(CHD~CAT+CHL+SMK,family=binomial,data=evans))
[1] 409.3424

Test for omitting both HPT and AGE from full model uses 409.3424 - 401.947 =
7.3954
Compared to a chi-squared on 2df
> 1-pchisq(7.3954,2)
[1] 0.02478046

> anova(glm(CHD~CAT+CHL+SMK,family=binomial,data=evans),
glm(CHD~CAT+AGE+CHL+SMK+HPT,family=binomial,data=evans),test="Chisq")
Analysis of Deviance Table

Model 1: CHD ~ CAT + CHL + SMK
Model 2: CHD ~ CAT + AGE + CHL + SMK + HPT
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       605     409.34
2       603     401.95  2   7.3955  0.02478 *
```

Interaction Terms

- We almost always observe a principle of hierarchy of models.
- If an interaction term such as CAT*HPT is in the model, then the main effects CAT and HPT are also in the model. If a three-way interaction such as CAT*CHL*HPT is in the model, then so are all three two way interactions CAT*CHL, CAT*HPT, and CHL*HPT as well as the three main effects.
- R will observe this in drop1() as long as the interactions are explicitly stated (CAT*CHL and not CC which has the same value).

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.132296   0.913911  -2.333 0.019640 *
CAT         -12.719878   3.138573  -4.053 5.06e-05 ***
CHL          -0.005312   0.004166  -1.275 0.202248          #Don't omit
SMK           0.698997   0.324996   2.151 0.031493 *
HPT           1.105883   0.328508   3.366 0.000762 ***
CAT:CHL       0.071175   0.014494   4.911 9.07e-07 ***
CAT:HPT      -2.221010   0.730937  -3.039 0.002377 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 352.92  on 602  degrees of freedom
AIC: 366.92

Number of Fisher Scoring iterations: 6
```

```
> drop1(glm(CHD~CAT+CHL+SMK+HPT+CAT*CHL+CAT*HPT,binomial,data=evans))
Single term deletions

Model:
CHD ~ CAT + CHL + SMK + HPT + CAT * CHL + CAT * HPT
        Df Deviance    AIC
<none>       352.92 366.92
SMK      1   357.93 369.93
CAT:CHL  1   399.88 411.88
CAT:HPT  1   362.06 374.06
```

Can’t drop CAT, CHL, or HPT.

Interaction Terms

- Other than the hierarchical model for interactions, we can compare any two nested models.
- If we want to omit CHL, we also have to omit CAT*CHL, and we have a 2df comparison.
- CHL is quantitative, CAT*CHL is 0 whenever CAT = 0, and is equal to CHL when CAT = 1.
- Inclusion of the interaction means that the effect of high catecholamines is different (greater) when cholesterol is high.
- It also means that the effect of cholesterol is different (greater) when catecholamines are high.