Logistic Regression

as a

Classification Tool

Logistic Regression as a Classification Tool

Logistic regression models may perform better than linear or quadratic discriminant functions when

- Some measured traits are categorical (e.g. male or female)
- Some measured traits are non-normal

Logistic Regression

For two populations the basic setup includes training samples from the two populations with

- Sample sizes n_1 and n_2 and $n = n_1 + n_2$
- Vectors of measured traits and the population indicator

$$(x_{1i},\ x_{2i},\ \ldots,\ x_{pi},\ Y_i) \qquad \text{for } j = 1, 2, ..., n$$

where

$$Y_i = \begin{cases} 1 & \text{for an observation from population 1} \\ 2 & \text{for an observation from population 2} \end{cases}$$

Logistic Regression

Construct a model to estimate the probability that the i-th case, with values $\underset{\sim}{\mathbf{x}}_i = (x_{1i}, x_{21}, \ldots .x_{pi})'$, came from population 1, i.e.,

$$\pi_i = P(Y_i = 1 \mid (x_{1i},\ x_{2i},\ \ldots,\ x_{pi}))$$

and the probability it came from population 2. i.e.,

$$1 - \pi_i = P(Y_i = 2 \mid (x_{1i},\ x_{2i},\ \ldots,\ x_{pi}))$$

Logistic regression relates the log-odds that the i-th case was sampled from population 1 to a linear function of some parameters, i.e.,

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

Logistic Regression

In the model

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

The coefficient β_2 is a partial regression coefficient representing the change in the log-odds (or logit) when

- x_2 is increased by one unit, and
- the values of x_1 and x_3 are held constant.

Also

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})}{1 + exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})}$$

Logistic Regression

- Several binary variables may be included in the model to represent the levels of a single classification factor
- Model the log-odds that a randomly selected adult would favor a proposal to raise property taxes to build a new elementary school with respect to some characteristic of the children associated the household:

  1. youngest child less than 5 years old
  2. youngest child less than 12 years old
  3. youngest child at least 12 years old
  4. no children

Logistic Regression

Define a set of four "dummy" variables:

$$x_{1i} = \begin{cases} 1 & \text{if the youngest child in the i-th household is under 5} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{if the youngest child in the i-th household is under 12} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{3i} = \begin{cases} 1 & \text{if the youngest child in the i-th household is at least 12} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{4i} = \begin{cases} 1 & \text{if the i-th household has no children} \\ 0 & \text{otherwise} \end{cases}$$

Logistic Regression

A logistic regression model is

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$$

This model could be specified in R with a factor, but it may be advantageous to use individual binary explanatory variables for searching for a good model.

- Stepwise search procedures will keep or drop an entire factor (use all of the categories or use none).
- By using individual binary explanatory variables it is possible to keep just two categories, or any subset of the categories.

Logistic Regression

- There are two many parameters in this model. There are four logits for the four different categories and five parameters β_0, β_1, β_2, β_3, β_4

- You must eliminate one of the parameters. You can impose the restriction $\beta_4 = 0$, and the resulting model is

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

- This gives a particular meaning to the parameters.

Logistic Regression

- The odds of agreement for households with no children is

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0) = \beta_0$$

  and the probability of agreement for households with no children is

$$\pi = \frac{exp(\beta_0)}{1 + exp(\beta_0)}$$

- For households with a youngest child under five, the odds of agreement are

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \quad \text{and} \quad \pi = \frac{exp(\beta_0 + \beta_1)}{1 + exp(\beta_0 + \beta_1)}$$

Logistic Regression

β_k represents the natural logarithm of the odds ratio that a respondent randomly selected from the k-th type of household will agree, relative to the baseline category (households with no children).

$$
\begin{aligned}
\beta_k &= (\beta_0 + \beta_k) - \beta_0 \\
&= log\left(\frac{\pi_k}{1-\pi_k}\right) - \beta_0 \\
&= log\left(\frac{\pi_k}{1-\pi_k}\right) - log\left(\frac{\pi_4}{1-\pi_4}\right) = log\left(\frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_4}{1-\pi_4}}\right)
\end{aligned}
$$

and

$$
e^{\beta_k} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_4}{1-\pi_4}} \quad \text{is an odds ratio}
$$

Logistic Regression

Suppose

$$e^{\beta_k} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_4}{1-\pi_4}} = 4.$$

then the odds of agreement for a respondent randomly selected from the k-th type of household are four times greater than the odds of agreement for a respondent randomly selected from the baseline category (household with no children).

- The previous discussion was presented in terms of the population parameters which are unknown in practical situations.
- We will need to collect some data (called training samples) and estimate the parameters.
- We will use maximum likelihood estimation (assumes simple random sampling).

Rehabilitation Program

Story County, Iowa had a program for rehabilitating criminal offenders with certain addictions and behavioral problems.

- Objective: Some judges wanted help in discriminating between criminal offenders who could be helped by the program and those who would not be helped.

- Training samples:From past experience, the program was successful for 31 subjects and it was unsuccessful for 57 subjects that the judges had assigned to the program.

- For subjects previously enrolled in the program, information was available on the following variables

Rehabilitation Program

- AGE: Age of the subject in years
- SEX: (1=female, 2=male)
- EDUC: education level (1=elementary, 2=some high school, 3=high school graduate, 4=college graduate)
- EMOTION: score on a psychological test for emotional problems
- ETREAT: previous treatment for emotional problems (1=yes, 2=no)
- LIVING: living arrangement (1=alone, 2=with parents, 3=with friends 4=with spouse, 5=in an institution)

Rehabilitation Program

- ATREAT: previously treated for alcoholism (1=yes, 2=no)
- ALCADD: score on a test for alcohol related problems
- HEALTH: health problems (1=yes, 2=no)
- FINANCE: financial problems (1=yes, 2=no)
- Marriage: marital status (1=good relationship with spouse,
2=poor relationship with spouse
3=divorced or separated
4=single
- PDRINK parental drinking (1=yes, 2=no)

Rehabilitation Program

- SIBS: number of siblings
- WORK: currently employed (1=yes, 2=no)
- WAGES: Yearly wages in thousands of dollars
- JOBS: Number of jobs held in the past five years
- DAGE: Age when subject started using alcohol
- DFREQ: Number of days per week in which alcohol was consumed
- STOP: Did the subject try to stop drinking in the past (1=yes, 2=no)

Rehabilitation Program

- DRY: Longest period in which subject did not drink (in months)
- DRUGS: drug dependencies other than alcohol (1=yes, 2=no)

Discrimination with Logistic Regression

The judges would like to have a logistic regression formula

$$log\left(\frac{\pi_{success}}{1-\pi_{success}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

based on some or all of the 21 variables, to estimate the conditional odds (or the probability) that the program will be successful for specific future offenders that come before the court.

Note that

$$\pi_{success} = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

If this probability is high enough, a judge may refer the offender to the rehabilitation program.

Discrimination with Logistic Regression

A judge can make two types of correct decisions:

- Refer an offender who would be successful into the rehabilitation program.
- Do not refer an offender who would fail.

A judge could make two types of mistakes:

- Do not refer an offender who would be successful.
- Refer an offender who fails.

If costs can be assigned to these two types of mistakes, we can try to find the set of explanatory variables the minimizes the expected misclassification cost, or just minimize the total misclassification probability.

Discrimination with Logistic Regression

We need to use the information in the training samples to estimate the parameters in the model.

- Training sample 1: $n_1 = 31$ previous offenders who were successful with the program.
- Training sample 2: $n_2 = 57$ previous offenders who were not helped by the program.

We may not want to use all 21 variables:

- Only use variables that are good discriminators. Using too many variables that provide little information for discriminating between the "successful" and "failure" populations may reduce the accuracy of the classification procedure.
- Avoid variables with too many missing values.

Rehabilitation Program

```
#  This R code performs logistic regression on alcoholic
#  rehabilitation data posted as crimeR.dat.

#  Enter the data and assign variable names

crim<-read.table("crimeR.txt",
     header=F, col.names=c("ID","result","age","sex","educ","emotion",
     "etreat", "living","atreat","alcadd","health","finance","marriage",
      "pdrink","sibs","work","wages","jobs","dage","dfreq","stop",
      "dry","drugs"))

head(crim)
  ID result age sex educ emotion etreat living atreat alcadd health finance
1  1      1  37   1    4      14      1      2      1     29      1       1
2  2      1  30   1    3      74      2      4      1     71      1       1
3  3      1  26   1    2      50      2      4      2     30      2       1

  marriage pdrink sibs work wages jobs dage dfreq stop dry drugs
1        3      1    3    1     6    2   21     4    1  12     1
2        2      1    6    1    16    7   22     7    1   6     1
3        2      1    5    1    12   12   20     2    1  12     1
```

```
# Create binary variables from categorical variables.  This was done instead of
# using factors so individual levels could be selected in a model search.

   nn<-dim(crim)[1]
   crim$e1<-rep(0,nn)
   crim$e1[crim$educ==1]<-1
   crim$e2<-rep(0,nn)
   crim$e2[crim$educ==2]<-1
   crim$e3<-rep(0,nn)
   crim$e3[crim$educ==3]<-1
   crim$m1<-rep(0,nn)
   crim$m1[crim$marriage==1]<-1
   crim$m2<-rep(0,nn)
   crim$m2[crim$marriage==2]<-1
   crim$m3<-rep(0,nn)
   crim$m3[crim$marriage==3]<-1
   crim$L1<-rep(0,nn)
   crim$L1[crim$living==1]<-1
   crim$L2<-rep(0,nn)
   crim$L2[crim$living==2]<-1
   crim$L3<-rep(0,nn)
   crim$L3[crim$living==3]<-1
   crim$L4<-rep(0,nn)
   crim$L4[crim$living==4]<-1
```

```
# Transform the binary response to take values 0  for success and 1 for failure
  crim$result <- crim$result-1
  head(crim)

#  Fit a big logistic regression model.  Cases with incomplete
#  information will not be used.

    crim1<-glm(result ~ age+sex+e1+e2+e3+emotion+etreat+L1+L2+L3+L4+
                   atreat+alcadd+health+finance+m1+m2+m3+pdrink+
                   sibs+work+wages+jobs+dage+dfreq+stop+dry+drugs,
                   family=binomial, data=crim)

      Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred

   crim1$coef
  (Intercept)           age           sex            e1            e2
 2.520628e+15 -1.319168e+14  1.654343e+15  1.238782e+15 -1.793969e+15
           e3       emotion        etreat            L1            L2
-7.788557e+14  6.967789e+13 -1.563272e+15 -2.109283e+13 -6.885898e+14
           L3            L4        atreat        alcadd        health
 5.143032e+14 -2.278033e+15 -1.025632e+15 -2.010953e+13 -1.359668e+15
      finance            m1            m2            m3        pdrink
-1.003867e+15  5.706851e+15  3.190179e+15  3.662664e+14  1.390609e+15
         sibs          work         wages          jobs          dage
 4.833469e+13  1.397959e+14 -3.528736e+13 -8.820685e+12  1.542387e+14
        dfreq          stop           dry         drugs
-2.617661e+14 -9.074112e+13  3.576401e+12  8.698692e+14
```

```
#  Select rows of data frame with no missing data.  This must be done
#  to avoid errors in using the stepwise search algorithms.

  crimc <- na.omit(crim)

#  Some coefficients are infinite.  The algorithm did
#  not converge to a finite solution.  Fit a smaller model.

   crim1<-glm(result ~ age+sex+emotion+etreat+L1+L2+L3+L4+
                  atreat+alcadd+health+finance+m1+m2+m3+pdrink+
                  sibs+work+wages+jobs+dage+dfreq+stop+dry+drugs,
                  family=binomial, data=crimc)

#  Print the estimated coefficients

    crim1$coef
 (Intercept)          age          sex      emotion       etreat           L1
  8.67945977  -0.40966277   2.95932905   0.03151005  -5.00955682   1.39881277
          L2           L3           L4       atreat       alcadd       health
 -2.45080759   3.50573389  13.35545480  -2.52029519   0.07292397  -1.46815951
     finance           m1           m2           m3       pdrink         sibs
  0.52487532  -4.92862723 -10.02094534   2.64227757   2.91026239   0.58553489
        work        wages         jobs         dage        dfreq         stop
  1.41459251  -0.25422258  -0.16374896   0.19301432  -0.57425357  -1.30531152
         dry        drugs
 -0.01559539   2.19785166
```

```
#  Use a backward selection algorithm to select a good model and
#  print the coefficients for the final model

   crim2<-step(crim1, direction=c("backward"))

Start:  AIC=90.79
result ~ age + sex + emotion + etreat + L1 + L2 + L3 + L4 + atreat +
    alcadd + health + finance + m1 + m2 + m3 + pdrink + sibs +
    work + wages + jobs + dage + dfreq + stop + dry + drugs

          Df Deviance    AIC
- m1       1   38.794 88.794
- m2       1   38.797 88.797
- L4       1   38.855 88.855
- finance  1   38.874 88.874
- dry      1   38.881 88.881
- emotion  1   38.907 88.907
- stop     1   39.062 89.062
- L1       1   39.107 89.107
- alcadd   1   39.160 89.160
- jobs     1   39.355 89.355
```

```
- work     1   39.575 89.575
- L2       1   39.902 89.902
- dage     1   39.934 89.934
- health   1   39.958 89.958
- drugs    1   40.037 90.037
- sex      1   40.129 90.129
- m3       1   40.584 90.584
- pdrink   1   40.607 90.607
<none>         38.794 90.794
- wages    1   41.025 91.025
- L3       1   41.242 91.242
- dfreq    1   42.141 92.141
- atreat   1   42.595 92.595
- sibs     1   42.632 92.632
- etreat   1   46.715 96.715
- age      1   49.407 99.407
```

```
Step:  AIC=88.79
result ~ age + sex + emotion + etreat + L1 + L2 + L3 + L4 + atreat +
    alcadd + health + finance + m2 + m3 + pdrink + sibs + work +
    wages + jobs + dage + dfreq + stop + dry + drugs

          Df Deviance    AIC
- finance  1   38.875 86.875
- dry      1   38.881 86.881
- emotion  1   38.907 86.907
- stop     1   39.062 87.062
- L1       1   39.107 87.107
- alcadd   1   39.160 87.160
- m2       1   39.201 87.201
- jobs     1   39.355 87.355
- work     1   39.575 87.575
- L2       1   39.903 87.903
   .       .      .
   .       .      .
   .       .      .
- sibs     1   42.637 90.637
- etreat   1   46.716 94.716
- age      1   49.452 97.452
```

After many more steps the search ends as follows:

```
Step:  AIC=71.77
result ~ age + etreat + L2 + atreat + health + pdrink + sibs +
    wages + dfreq

         Df Deviance    AIC
<none>        51.771 71.771
- dfreq   1   54.801 72.801
- wages   1   54.882 72.882
- pdrink  1   55.044 73.044
- sibs    1   55.495 73.495
- health  1   55.521 73.521
- L2      1   58.140 76.140
- atreat  1   58.405 76.405
- etreat  1   60.577 78.577
- age     1   64.199 82.199

   crim2$coef
(Intercept)         age      etreat          L2      atreat      health
 15.6898058  -0.1805976  -2.3497667  -2.6252080  -2.0251175  -1.7933833
     pdrink        sibs       wages       dfreq
  1.4002977   0.3297428  -0.1501270  -0.3464049
```

```
#  Use a stepwise selection algorithm to select a good model and
#  print the coefficients for the final model

   crim3<-glm(result ~ age, family=binomial, data=crimc)

   crim4<-step(crim3, direction=c("both"),
             scope=list(upper= ~age+sex+emotion+etreat+L1+L2+L3+L4+
                    atreat+alcadd+health+finance+m1+m2+m3+pdrink+sibs+
                    work+wages+jobs+dage+dfreq+stop+dry+drugs,
                    lower = ~1),trace=F)

  crim4$coef

(Intercept)         age      etreat        sibs       wages      atreat
 10.0855259  -0.1606608  -2.4659444   0.4572570  -0.1452901  -2.1506078
         m1          L2         sex
  4.8484758  -1.6614421   2.1930063
```

```
#  Create a new data frame using only the variables or
#  factors you want to include in the logistic regression
#  model.

   x<-subset(crim, select=c("age","etreat","L2","atreat",
                    "health","pdrink","wages","dfreq","result"))
   crimc2 <- na.omit(x)
   nnm <- nrow(crimc2)

   crim6<-glm(result ~ age+etreat+L2+health+pdrink+
                    atreat+wages+dfreq,family=binomial, data=crimc2)
   crim6p<-predict(crim6)
   crim6class <- rep(0,nnm)
   crim6class[ crim6p>0 ]<-1
   table(crimc2$result, crim6class)

  crim6class
     0  1
  0 16 15
  1  7 45
```

```
#  Compute cross validation estimates of the misclassification
#  probabilities.  First source in the function posted in
#  the file crossval2 from the directory where you stored it
#  This function is a modified version of the original ‘‘crossval’’
#  code given by Efron.

 source("crossval2.R")

 y <- crimc2$result
 resultcv2<-crossval2(crimc2[ , -9],y)
 table(y,resultcv2)

     resultcv2
  y      0    1
  0      8   23
  1     15   37

#  Create a new data frame using only the variables or
#  factors you want to include in the second logistic
#  regression model from the stepwise search.

   x2<-subset(crim, select=c("age","etreat","sibs","wages",
                "atreat","m1","L2","sex","result"))
   crimc2 <- na.omit(x2)
   nnm <- nrow(crimc2)
```

```
crim7<-glm(result ~ age+etreat+sibs+wages+atreat+
                m1+L2+sex,family=binomial, data=crimc2)
crim7p<-predict(crim7)
crim7class <- rep(0,nnm)
crim7class[ crim7p>0 ]<-1
table(crimc2$result, crim7class)

 crim7class
   0  1
0  15 16
1   7 40
```

```
y <- crimc2$result
 resultcv2<-crossval2(crimc2[ , -9],y)
 table(y,resultcv2)

 resultcv2
y    0  1
0   11 20
1   13 34
```

```
# Evaluate the model that used all of the variables
# More cases deleted because of more missing values.

 crim1<-glm(result ~ age+sex+emotion+etreat+L1+L2+L3+L4+
                  atreat+alcadd+health+finance+m1+m2+m3+pdrink+
                  sibs+work+wages+jobs+dage+dfreq+stop+dry+drugs,
                  family=binomial, data=crimc)

 nnm <- nrow(crimc)
 crim1p<-predict(crim1)
 crim1class <- rep(0,nnm)
 crim1class[ crim1p>0 ]<-1
 table(crimc$result, crim1class)

 crim1class
    0  1
0  22  5
1   5 29
```

Discrimination with Logistic Regression

Comments:

- We had some difficulty finding a good logistic regression model.
- Most of the variables did not provide much information for separating "successful" offenders from "unsuccessful" offenders. Better explanatory variables are needed.
- Training samples were small.
- Crossvalidation gave more honest estimates of misclassification probabilities.
- With larger training samples you could use a set aside method to assess the accuracy of the model.