BIOSTAT/EPI 536 Homework 4

Solution Key
October 25, 2007

In an attempt to find clinical correlates of ankle fracture, all eligible patients with ankle trauma at a particular clinic between September 1979 and November 1981 received physical examinations and ankle x-rays were taken. A Stata dataset associated with this study is available on the class website (ankle_full.dta). Three categories of ankle injury are considered: sprain, rupture and fracture. For the purposes of this analysis, consider the outcome variable FINALDX recoded to fracture versus non-fracture. (Be careful of missing values!) The data are divided into two samples (by the variable SAMPLE), an estimation sample (coded 0) and a validation sample (coded 1). For this exercise you will be using the following variables: fracture, AGE, SEX, PAP, NUMB, SWANK, PSPR1, PSPR2, PSPR3, AKTIV, MECH, SNAP, GIVWA, BONE1, BONE2, COLOR. Create a data set in which none of these variables have any missing values. For the purposes of this analysis use the following coding for AGE

Age = 0 if AGE<40
= (AGE-40) if AGE ≥ 40.

Use all other variables as they are coded.

COMMENTS:

- In order to replicate the exact numbers that we show in the solution, you need to delete records with missing values for the list of variables to be considered. You need to make comparisons across a common data set, not data sets that differ because of different constellations of missing values.

- In addition, the original model fitting should be carried out using the *estimation data set* (SAMPLE==0), N=395. Then the *validation set* should be used to test the adequacy (accuracy) of the model as it has been specified by the estimation data (so, using values of the coefficient estimates based on the model fitted to the estimation data) using new data (SAMPLE==1), N=180. You DO NOT generate new coefficients by rerunning the model on the validation data prior to looking at the goodness of fit, sensitivity/specificity, and ROC.

1. Using the estimation sample only*: Use forward and backward stepwise procedures to choose a subset of the above variables to predict ankle fracture.

SOLUTION: The forward selection procedure with entry and removal probabilities .2 and .25 respectively selects Age, COLOR, MECH, BONE1, PSPR1, PSPR2, GIVWA.

The backward elimination procedure with the same entry and removal probabilities selects Age, COLOR, MECH, BONE1, BONE2, PSPR1, PSPR3, GIVWA.

(STATA code and output for this question are provided at the end of the solution key.)

2. Using the estimation sample only*: Evaluate, in terms of their AIC, models containing the following variables:

(i) Age
(ii) Age, COLOR
(iii) Age, COLOR, BONE1
(iv) NUMB, SEX, AKTIV
(v) Age, COLOR, BONE1, PSPR1, MECH
(vi) PAP, NUMB, SNAP, SEX, SWANK
(vii) Age, COLOR, BONE1, PSPR1, MECH, PSPR2, GIVWA
(viii) PAP, NUMB, SNAP, PSPR1, MECH, PSPR2, BONE2
(ix) Age, COLOR, BONE1, PSPR1, MECH, PSPR3, GIVEWA, BONE2
(x) The models chosen in 1 above, if not already assessed.

Make a choice as to which of the above you would choose to model the probability of fracture and motivate your choice. Interpret the coefficients of the chosen model.

SOLUTION: The table below summarizes the 9 models in terms of their likelihood and AIC:

Variable	p	-logL	AIC
(i)	2	90.35	184.7
(ii)	3	85.49	177.0
(iii)	4	81.60	171.2
(iv)	4	103.52	215.0
(v)	6	73.16	158.3
(vi)	6	104.82	221.6
(vii)	8	70.84	157.7
(viii)	8	96.18	208.4
(ix)	9	70.12	158.2

Model (v) would be a parsimonious choice with low AIC. The forward (vii) and backward (ix) selection models also represent models with low AIC. The fitted model (v) is of the form:

logit(p)= -2.56+.103 Age+1.78 COLOR+1.42 BONE1-1.06 PSPR1-1.52 MECH

The model suggests that the odds of fracture increases with increasing age over 40 ($\hat{\psi}$ =2.8 for a 10 year increase in age among those over 40, other factors being equal) and is raised in individuals with discoloration and tender ankle bones. The odds of fracture decreases among those who have had previous sprains and among those where inversion caused the injury.

3. Evaluate the predictive accuracy of the model, using

(i) the estimation sample and
(ii) the validation sample.

Comment on results in (ii) and their comparison with (i).

SOLUTION:

(i) Because Age (among those over 40) is a continuous variable, there are 83 different covariate patterns with 395 observations. To evaluate overall goodness of fit, we need to group the data:

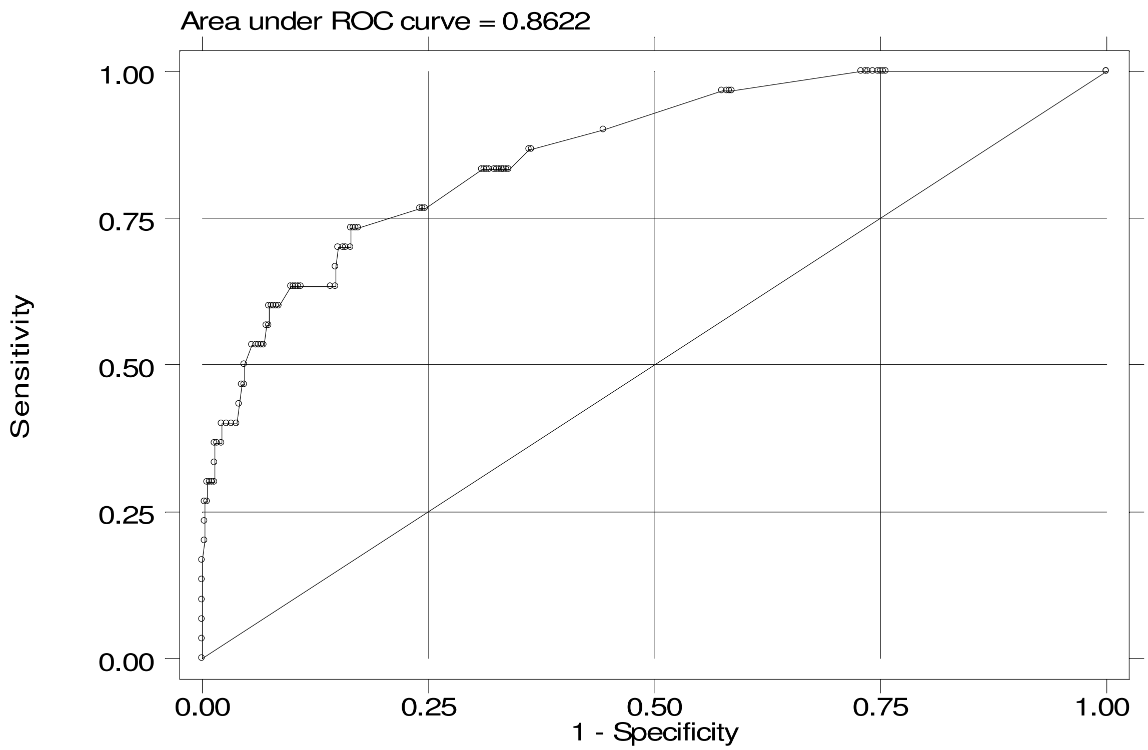
```
. lfit, group(10) table

Logistic model for fracture, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

Note: Because of ties, there are only 8 distinct quantiles.

_Group     _Prob     _Obs_1     _Exp_1     _Obs_0     _Exp_0     _Total
     2    0.0058          0        0.5         89       88.5         89
     3    0.0166          1        1.0         62       62.0         63
     5    0.0235          2        1.3         52       52.7         54
     6    0.0325          1        0.8         30       30.2         31
     7    0.0651          3        2.5         42       42.5         45
     8    0.0865          2        2.5         32       31.5         34
     9    0.2013          5        4.7         35       35.3         40
    10    0.9600         16       16.7         23       22.3         39

       number of observations =       395
             number of groups =         8
      Hosmer-Lemeshow chi2(6) =         1.30
                  Prob > chi2 =         0.9716
```

The Hosmer-Lemeshow test shows no evidence of lack of fit (note, the goodness of fit test is optional here, as the question asks only that you evaluate “predictive accuracy”).

```
. estat classification

Logistic model for fracture

              -------- True --------
Classified |         D            ~D         Total
-----------+--------------------------+-----------
     +     |         8             2  |         10
     -     |        22           363  |        385
-----------+--------------------------+-----------
   Total   |        30           365  |        395

Classified + if predicted Pr(D) >= .5
True D defined as fracture ~= 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   26.67%
Specificity                     Pr( -|~D)   99.45%
Positive predictive value       Pr( D| +)   80.00%
Negative predictive value       Pr(~D| -)   94.29%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.55%
False - rate for true D         Pr( -| D)   73.33%
False + rate for classified +   Pr(~D| +)   20.00%
False - rate for classified -   Pr( D| -)    5.71%
--------------------------------------------------
Correctly classified                        93.92%
--------------------------------------------------
```

At a threshold of .5 (i.e. all individuals with estimated probability of fracture > .5 are classified as “fracture”), the estimated sensitivity and specificity are 27% and 99% respectively. Decreasing the threshold will increase the sensitivity but decrease the specificity.

```
. lroc

Logistic model for fracture

number of observations =      395
area under ROC curve   =   0.8622
```



The ROC curve displays the trade-off between sensitivity and specificity for varying probability classification thresholds. The choice of threshold (if any) would depend on the consequences of false positives and false negatives.

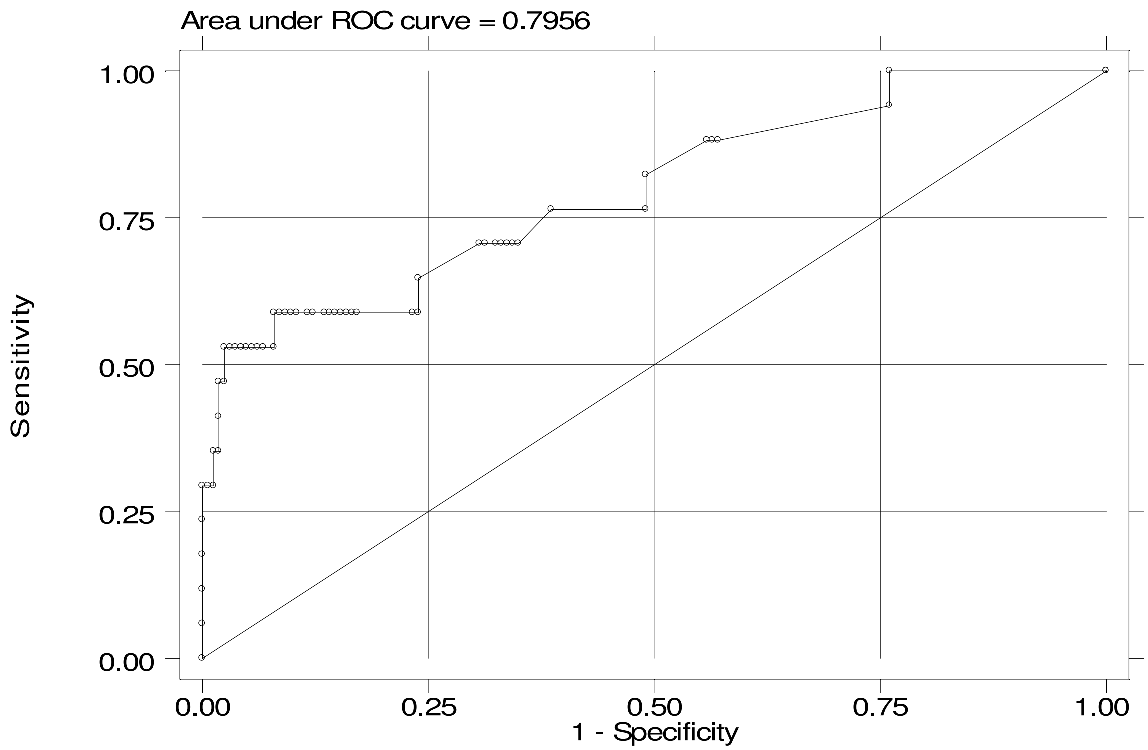
(ii) In evaluating the fit for the second sample by dividing the data into 8 groups, we should have 8 degrees of freedom as we are not using the validation data for estimation (in fact, because of ties, there are only 7 distinct groups/degrees of freedom). There is no evidence that the model developed on the estimation sample is not a good fit to the validation data set.

```
. estat gof if SAMPLE==1, group(8) table out

Logistic model for fracture, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  (There are only 7 distinct quantiles because of ties)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0079 |     0 |   0.3 |    39 |  38.7 |    39 |
  |     3 | 0.0180 |     1 |   0.6 |    31 |  31.4 |    32 |
  |     4 | 0.0220 |     1 |   0.4 |    18 |  18.6 |    19 |
  |     5 | 0.0487 |     3 |   1.3 |    29 |  30.7 |    32 |
  |     6 | 0.0838 |     1 |   1.4 |    17 |  16.6 |    18 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     7 | 0.1312 |     1 |   1.9 |    17 |  16.1 |    18 |
  |     8 | 0.9117 |    10 |   9.2 |    12 |  12.8 |    22 |
  +--------------------------------------------------------+
```

```
       number of observations =       180
             number of groups =         7
      Hosmer-Lemeshow chi2(7) =         4.33
                  Prob > chi2 =         0.7412


. estat classification if SAMPLE==1

Logistic model for fracture

              -------- True --------
Classified |         D            ~D         Total
-----------+--------------------------+-----------
     +     |         6             2  |          8
     -     |        11           161  |        172
-----------+--------------------------+-----------
   Total   |        17           163  |        180

Classified + if predicted Pr(D) >= .5
True D defined as fracture ~= 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   35.29%
Specificity                     Pr( -|~D)   98.77%
Positive predictive value       Pr( D| +)   75.00%
Negative predictive value       Pr(~D| -)   93.60%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.23%
False - rate for true D         Pr( -| D)   64.71%
False + rate for classified +   Pr(~D| +)   25.00%
False - rate for classified -   Pr( D| -)    6.40%
--------------------------------------------------
Correctly classified                        92.78%
--------------------------------------------------
```

Based on the validation sample, and a threshold of .5, the overall accuracy is about the same as that estimated using the estimation sample. The ROC below has slightly smaller area, but represents comparable performance.

```
. lroc if SAMPLE==1

Logistic model for fracture

number of observations =      180
area under ROC curve   =   0.7956
```



Area under ROC curve = 0.7956
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 - Specificity

STATA CODE AND OUTPUT, question 1:

```
. gen age2=( AGE>40)*(AGE-40)
. sw logistic fracture age2 SEX PAP NUMB SWANK PSPR1 PSPR2 PSPR3 AKTIV MECH
  SNAP GIVWA BONE1 BONE2 COLOR if SAMPLE==0, forw pe(.2) pr(.25)

                      begin with empty model
p = 0.0000 <  0.2000  adding   age2
p = 0.0008 <  0.2000  adding   COLOR
p = 0.0023 <  0.2000  adding   MECH
p = 0.0022 <  0.2000  adding   BONE1
p = 0.0201 <  0.2000  adding   PSPR1
p = 0.1111 <  0.2000  adding   GIVWA
p = 0.1568 <  0.2000  adding   PSPR2

Logit estimates                                   Number of obs   =        395
                                                  LR chi2(7)      =      70.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -70.839618                       Pseudo R2       =     0.3327

------------------------------------------------------------------------------
fracture | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    age2 |   1.116705    .025077      4.915   0.000       1.068621    1.166953
   COLOR |   7.520955   4.722962      3.213   0.001       2.196548    25.75166
    MECH |   .2114585   .1007782     -3.260   0.001        .083091    .5381408
   BONE1 |   4.918335   2.395903      3.270   0.001       1.893069     12.7782
   PSPR1 |   .2151405   .1315445     -2.513   0.012       .0649034    .7131432
   GIVWA |   .4372808    .218297     -1.657   0.098       .1643721    1.163303
   PSPR2 |    1.81017     .75866      1.416   0.157       .7961088    4.115914
------------------------------------------------------------------------------


. sw logistic fracture age2 SEX PAP NUMB SWANK PSPR1 PSPR2 PSPR3 AKTIV MECH
  SNAP GIVWA BONE1 BONE2 COLOR if SAMPLE==0, pe(.2) pr(.25)

                      begin with full model
p = 0.7156 >= 0.2500  removing SEX
p = 0.6342 >= 0.2500  removing NUMB
p = 0.7035 >= 0.2500  removing PAP
p = 0.5378 >= 0.2500  removing AKTIV
p = 0.5145 >= 0.2500  removing SWANK
p = 0.3275 >= 0.2500  removing SNAP
p = 0.2944 >= 0.2500  removing PSPR2

Logit estimates                                   Number of obs   =        395
                                                  LR chi2(8)      =      72.08
                                                  Prob > chi2     =     0.0000
Log likelihood = -70.121817                       Pseudo R2       =     0.3395

------------------------------------------------------------------------------
fracture | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    age2 |    1.11273   .0244579      4.860   0.000       1.065811    1.161714
   COLOR |   6.476592   3.928618      3.080   0.002       1.972507    21.26545
   BONE1 |   4.407416   2.121786      3.081   0.002       1.715556    11.32305
   BONE2 |   1.788628   .8550446      1.216   0.224       .7008216    4.564914
    MECH |   .1989013   .0961001     -3.343   0.001       .0771571    .5127428
   PSPR1 |   .1885994   .1246831     -2.523   0.012       .0516192    .6890803
   GIVWA |   .4104049   .2063274     -1.772   0.076        .153207    1.099377
   PSPR3 |   1.964432   .9012763      1.472   0.141       .7992964    4.827986
------------------------------------------------------------------------------
```