BIOST/EPI 536
Autumn Quarter 2007

HOMEWORK 4
Due Thursday 25th October

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 $\geq$ 40.

Use all other variables as they are coded.

Using the estimation sample only*:

1. Use forward and backward stepwise procedures to choose a subset of the above variables to predict ankle fracture.

2. 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.

* include "if SAMPLE==0" at the end of each Stata logistic model fitting command. See the Stata manual for further information.

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).