Homework Solutions
Applied Logistic Regression

WEEK 8

Exercise 2:

Use the hyponatremia.dta dataset.

(a) Fit a regression model with **female**, **urinat3p**, **runtime**, **wtdiff** and **bmi**. Evaluate the fit of the model using both Hosmer-Lemeshow test (using deciles of risk) and the Pearson Chi-square statistic. Assess whether the results of the two tests are consistent.

The "estat gof" command computes the Pearson's Chi-squared goodness of fit test whereas the command "estat gof, group(10) table" computes the Hosmer-Lemeshow goodness of fit test of the latest model.

```
. estat gof, group(10) table

Logistic model for nas135, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0078 |     0 |   0.2 |    45 |  44.8 |    45 |
  |     2 | 0.0151 |     1 |   0.5 |    43 |  43.5 |    44 |
  |     3 | 0.0241 |     0 |   0.9 |    44 |  43.1 |    44 |
  |     4 | 0.0374 |     0 |   1.3 |    44 |  42.7 |    44 |
  |     5 | 0.0552 |     4 |   2.0 |    40 |  42.0 |    44 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.0876 |     4 |   3.2 |    41 |  41.8 |    45 |
  |     7 | 0.1377 |     5 |   4.8 |    39 |  39.2 |    44 |
  |     8 | 0.2117 |     8 |   7.4 |    36 |  36.6 |    44 |
  |     9 | 0.3368 |    10 |  11.7 |    34 |  32.3 |    44 |
  |    10 | 0.9004 |    22 |  22.0 |    22 |  22.0 |    44 |
  +--------------------------------------------------------+

      number of observations =       442
            number of groups =        10
     Hosmer-Lemeshow chi2(8) =      5.75
                Prob > chi2 =      0.6755

. estat gof

Logistic model for nas135, goodness-of-fit test

      number of observations =       442
 number of covariate patterns =       441
            Pearson chi2(436) =       384.30
                 Prob > chi2 =         0.9642
```

Both tests indicate a good overall fit of the model. P-values differ because the number of degrees of freedom widely differs between the 2 tests.

(b) On the basis of the logistic model with **runtime**, **bmi**, **bmi2** and **wtdiff** as covariates, estimate the sensitivity and specificity of classifying subjects as having or not having hyponatremia using the cut-off values for the probability of hyponatremia of 0.5.

The "estat class, cut(0.5)" command gives the classification of the set of observations at a cutoff value of 0.5.

```
. estat class, cut(.5)

Logistic model for nas135

                -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        11              6  |         17
     -     |        43            382  |        425
-----------+--------------------------+-----------
   Total   |        54            388  |        442

Classified + if predicted Pr(D) >= .5
True D defined as nas135 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   20.37%
Specificity                     Pr( -|~D)   98.45%
Positive predictive value       Pr( D| +)   64.71%
Negative predictive value       Pr(~D| -)   89.88%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.55%
False - rate for true D         Pr( -| D)   79.63%
False + rate for classified +   Pr(~D| +)   35.29%
False - rate for classified -   Pr( D| -)   10.12%
--------------------------------------------------
Correctly classified                        88.91%
--------------------------------------------------
```

If the cut-off is 0.5, then sensitivity is 0.20, specificity is 0.98.

(c) Repeat the previous exercise using different cut-off values for the probability of hyponatremia. Draw by hand the ROC curve using these values of sensitivity and specificity.

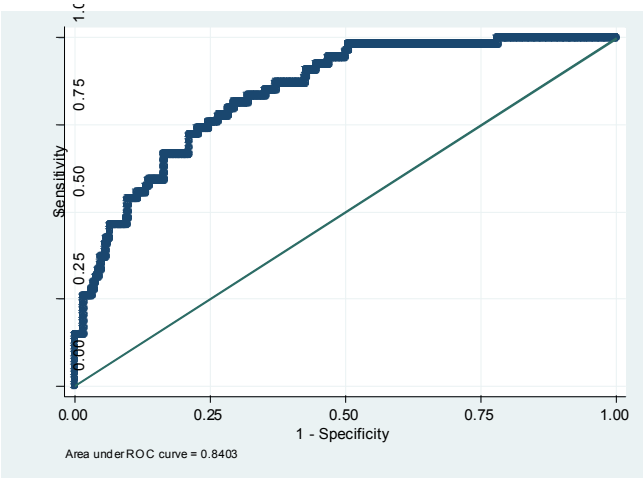
Cut-off	*Sensitivity*	*Specificity*
0	1	0
0.01	1	0.142
0.05	0.926	0.544
0.1	0.815	0.699
0.2	0.593	0.845
0.3	0.463	0.925
0.4	0.278	0.964
0.6	0.148	0.997
0.8	0.056	1
1	0	1

(d) Use Stata to obtain the ROC curve. What is the discriminatory power of the model?

```
. lroc

Logistic model for nas135

number of observations =       442
area under ROC curve   =    0.8403
```



The ROC curve shows excellent discrimination, with AUC=0.84

(e) Suppose someone had fraudulently access to your PC and altered data of the dependent variable in such a way that the coefficients of the model would remain the same. However, the predicted probabilities of the outcome would be largely affected. What would happen to goodness-of-fit statistics?

The overall fit of the model would be poorer. The discrimination of the model, as expressed by the ROC curve, would remain the same.

(f) Fit a model with **female** and **urinat3p** as covariates. Assess the overall fit of the model and its discriminatory power by computing the Pearson Chi square goodness of fit statistic and the area under the ROC curve.

Type "logit nas135 female urinat3p" in the command window to fit the logistic regression output. Then type "estat gof" to assess the fit of the model. The "lroc" command can also be used to obtain the ROC curve.

```
. estat gof

Logistic model for nas135, goodness-of-fit test

       number of observations =       480
 number of covariate patterns =         4
              Pearson chi2(1) =      0.26
                  Prob > chi2 =    0.6070

. lroc

Logistic model for nas135

number of observations =      480
area under ROC curve   =   0.6617
```

The p-value for the goodness-of-fit statistic is 0.607, whereas the AUC curve is 0.66

(g) Estimate the predicted probability of the outcome.

```
. predict yhat
(option pr assumed; Pr(nas135))
(8 missing values generated)

. tab yhat

 Pr(nas135) |      Freq.     Percent        Cum.
------------+-----------------------------------
   .0684694 |        299       62.29       62.29
   .1982314 |        154       32.08       94.38
   .2263822 |         20        4.17       98.54
    .496051 |          7        1.46      100.00
------------+-----------------------------------
      Total |        480      100.00
```