Homework Solutions
Applied Logistic Regression

WEEK 4

Exercise 1:

The data in hyponatremia.dta derive from an epidemiological study of hyponatremia (a life-threatening condition) among runners of the 2002 Boston Marathon. Hyponatremia is defined as an electrolyte disturbance in which the serum sodium concentration is lower than normal (<135 mmol/l). The aim of the study was to determine whether a runner experienced hyponatremia and to identify the principal risk factors. Participants in the 2002 Boston Marathon completed a survey including demographic and anthropometric characteristics (Body Mass Index) one or two days before the race. After the race, runners provided a blood sample in order to measure their serum sodium concentration and completed a questionnaire detailing their urine output during the race. Prerace and postrace weights were also recorded.

a. Perform a logistic regression analysis with Stata using **nas135** as dependent variable and **female** as the only independent variable. Interpret the coefficients of the model.

To perform the logistic regression analysis, type "logit nas135 female" into the command window.

```
. logit nas135 female, nolog

Logistic regression                               Number of obs   =        488
                                                  LR chi2(1)      =      19.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -175.96547                       Pseudo R2       =     0.0529

------------------------------------------------------------------------------
      nas135 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.225962    .279546     4.39   0.000     .6780617    1.773862
       _cons |  -2.474856   .2082475   -11.88   0.000    -2.883014   -2.066699
------------------------------------------------------------------------------

. di exp(-2.474856)
.08417511

. di 25/297
.08417508
```

The coefficient for female indicates the log odds ratio of hyponatremia of a female compared to a male. The constant indicates the log odds of hyponatremia for male. In this case, exp(1.225962)= 3.047, meaning that the odds of hyponatremia among females is 3.05 times that of males.

b. Fit a model with **runtime** as the only independent variable. Interpret the coefficient for **runtime**.

To perform the logistic regression analysis, type "logit nas135 runtime" into the command window.

```
. logit nas135 runtime, nolog

Logistic regression                                     Number of obs   =        477
                                                        LR chi2(1)      =      25.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -167.77184                             Pseudo R2       =     0.0702

------------------------------------------------------------------------------
      nas135 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     runtime |   .0155019   .0030909     5.02   0.000     .0094439    .0215599
       _cons |  -5.592594    .771282    -7.25   0.000    -7.104278   -4.080909
------------------------------------------------------------------------------
```

The coefficient for runtime indicates that the log odds of hyponatremia increases by 0.0155 each minute the marathon takes to be completed.

c. Calculate the Odds Ratio for the variable **runtime** and interpret it.

To generate the odds ratio instead of the logit coefficients, type "logistic nas135 runtime" into the command window.

```
. logistic nas135 runtime, nolog

Logistic regression                                     Number of obs   =        477
                                                        LR chi2(1)      =      25.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -167.77184                             Pseudo R2       =     0.0702

------------------------------------------------------------------------------
      nas135 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     runtime |   1.015623   .0031392     5.02   0.000     1.009489    1.021794
------------------------------------------------------------------------------
```

The ratio of the odds of hyponatremia of a runner who is 1 minute slower than another runner is 1.0156. This indicates that the odds of hyponatremia increases by 1.56% for each additional minute it takes to complete the marathon. This ratio is constant no matter what values of runtime are compared, provided that the difference in time is 1 minute (eg. 201 minutes vs 200 minutes, 251 minutes vs 250 minutes, etc.).

d. Interpret the coefficient for the constant in the model with **runtime** as the only independent variable. Does it make sense? If not, what can you do to obtain a coefficient for the constant which is easily interpreted?

This coefficient gives the log odds of hyponatremia of a runner who takes 0 minutes to complete the marathon. It does not make any sense! If runtime were centered around the mean (i.e. if a new variable called runtime2 were created by subtracting the mean value of runtime to all observations) then the coefficient would have indicated the log odds of hyponatremia of a runner who completes the marathon in the average time.

e. Calculate the Odds Ratio of hyponatremia of a runner who takes 2 hours more than another runner, and the corresponding 95% Confidence Interval

For this problem, remember to multiply both the upper and lower bounds of the confidence interval (found in the output in part c) by 120 minutes (2 hours).

```
. di exp(0.0155*120)
6.4237368

. di exp(0.0094439*120)
3.1057897

. di exp(0.0215599*120)
13.292341
```

The odds of hyponatremia of a runner who takes 2 hours more than another runner is about 6 times larger. This ratio can be as low as 3 times and as high as 13 times.

f. Fit a model with **female** and **runtime** as independent variables. Interpret both coefficients.

To perform this multiple logistic regression analysis, type "logit nas135 female runtime" into the command window.

```
. logit nas135 female runtime, nolog

Logistic regression                               Number of obs   =        477
                                                  LR chi2(2)      =      36.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -162.23985                       Pseudo R2       =     0.1009

------------------------------------------------------------------------------
      nas135 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .9638364    .291049     3.31   0.001     .3933908    1.534282
     runtime |   .0142136   .0032947     4.31   0.000     .0077562     .020671
       _cons |  -5.721056    .823284    -6.95   0.000    -7.334663   -4.107449
------------------------------------------------------------------------------
```

The coefficient for female is the log odds of hyponatremia of a female compared to a male who completes the marathon in the same time. The coefficient for runtime is the log odds of hyponatremia for an additional minute that it takes to complete the marathon, independent from the fact that the runner is a male or a female.

g. Compare the coefficients for **female** in the model with **female** as the only independent variable with that in the model that contains **female** and **runtime**. What is the percentage change in the coefficient of **female**?

```
. di (1.226-
0.964)*100/1.226
21.37031
```

There is a 21.4% change in the coefficient for female. This suggests possible confounding by runtime, provided that there is no interaction