Homework Solutions
Applied Logistic Regression

WEEK 5

Use the hyponatremia.dta dataset.

a. Assess the association between hyponatremia (dichotomous variable **nas135**) and sex (variable **female**) by making a 2 by 2 table. Calculate the odds ratio of hyponatremia of a female compared to a male. Compute the 95% confidence interval for this odds ratio. Interpret the findings.

You can calculate the odds, odds ratio, variance and subsequently the confidence interval by generating a 2X2 contingency table of gender by hyponatremia (using the “tab” command). Hand calculate the confidence interval for the odds ratio (or use the display “di” command), using the following formulas:

	Disease	No Disease
Exposure	A	B
No Exposure	C	D

Odds Ratio= $\frac{AD}{BC}$

Standard error= $\sqrt{Variance}$= $\sqrt{(\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D})}$

Note: When you calculate the upper and lower bounds of the odds ratio, make sure you first take the natural log of the OR and enter it into the formula for the confidence interval (the standard error is in the logit form). Afterwards, exponentiate both sides to convert it from the CI of the logit to the CI of the OR.

```
. tab female nas135

           |     Serum sodium
           | concentration <= 135
           |      mmol/liter
    Female |         0          1 |     Total
-----------+----------------------+----------
        No |       297         25 |       322
       Yes |       129         37 |       166
-----------+----------------------+----------
     Total |       426         62 |       488


. di (297*37)/(129*25)
3.4074419

. di ln((297*37)/(129*25))
1.2259618
```

```
. di sqrt(1/297+1/25+1/129+1/37)
.279546

. di 1.226-1.96*0.2796
.677984

. di exp(0.678)
1.9699339

. di 1.226+1.96*0.2796
1.774016

. di exp(1.774)
5.8943838
```

The odds of a female experiencing hyponatremia is 3.4 times greater than that of a male. The 95% Confidence interval for the odds ratio is (1.97, 5.89). Upon repeated sampling, 95% of confidence intervals constructed this way would cover the true population odds ratio.

b. Perform a logistic regression analysis with Stata using **nas135** as dependent variable and **female** as the only independent variable. Use the Likelihood Ratio test to assess the significance of the model. Is the model with **female** a better model than the naïve model? Use the Stata's built-in statistical functions to obtain p-values (type *help* functions)

```
. logit nas135 female

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
------------------------------------------------------------------------------
```

H_0: $\beta_{female}=0$ H_a: $\beta_{female}\neq0$

The LR test is given in the computer output. The chi-square statistic is 19.67 with 1 degree of freedom, which yields a p-value<0.0001. The model with female is significantly better than the naïve model.

You can also calculate the test statistic by hand through the following formula:

$$D = -2\ln(likelihood\ null\ model) + 2\ln\ (likelihood\ alternative\ model)$$

```
. di 2*(185.8-175.9655)
19.669

. di chi2tail(1,19.67)
9.203e-06
```

c. What is the naïve model? What is the probability of hyponatremia that this model predict?

The naïve model is the model with no independent variables. The “tab” command will give you the frequency and percent of hyponatremia in the dataset.

```
. tab nas135

      Serum |
     sodium |
concentrati |
  on <= 135 |
 mmol/liter |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        426       87.30       87.30
          1 |         62       12.70      100.00
------------+-----------------------------------
      Total |        488      100.00
```

The naïve model predicts a 12.7% probability of hyponatremia for every subject in the study.

d. Run a logistic regression analyses with no independent variables. Transform the coefficient obtained from this model into a probability.

First, run the naïve model through by typing the “logit” command, followed by the dependent variable (in this case “nas135).

The odds can be calculated by exponentiating the coefficient for the constant term, and probability can then be calculated from the odds through the following formula:

$$Probability = \frac{Odds}{1 + Odds}$$

```
. logit nas135

Iteration 0:   log likelihood = -185.80042
Iteration 1:   log likelihood = -185.80042

Logistic regression                               Number of obs   =        488
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -185.80042                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
      nas135 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -1.927305   .1359281   -14.18   0.000    -2.193719   -1.660891
------------------------------------------------------------------------------

. di exp(-1.927)/(1+exp(-1.927))
.12708301
```

Alternatively, after running the naïve model, Stata *predict* command gives the predicted probability of the outcome.

```
. predict p0
(option pr assumed; Pr(nas135))

. tab p0

 Pr(nas135) |      Freq.     Percent        Cum.
------------+-----------------------------------
   .1270492 |        488      100.00      100.00
------------+-----------------------------------
      Total |        488      100.00
```

e. Using the model with **female** as independent variable, compute the estimated probability of hyponatremia per males and females. Write down the equation for the logit.

To compute the estimated probability of hyponatremia by sex, first run the logit (note: the "quietly" command below suppresses the output of fitting the regression). Next, type "predict yhat" into the command box to generate a new variable of the estimated probability of hyponatremia given the model. By typing "tab yhat" into the command window, you can then see the two probabilities of hyponatremia by sex (as sex is the only independent variable in this model).

```
. quietly logit nas135 female

. predict yhat
(option pr assumed; Pr(nas135))

. tab yhat

 Pr(nas135) |      Freq.     Percent        Cum.
------------+-----------------------------------
   .0776398 |        322       65.98       65.98
   .2228916 |        166       34.02      100.00
------------+-----------------------------------
      Total |        488      100.00
```

The probability for a randomly chosen male is 7.8%, and that for a randomly chosen female is 22.3%. The equation for the logit is $g(x) = \beta_{0+}\beta_1[female]$

$$g(x) = \beta_0 + \beta_1[female] = -2.4749 + 1.2260[female]$$