Homework Solutions
Applied Logistic Regression

WEEK 1

Exercise 1:

Use the Myopia Study (MYOPIA.dta)

One variable that is clearly important is the initial value of spherical equivalent refraction (SPHEQ).

a. Write down the equation for the logistic regression model of SPHEQ on MYOPIA. Write down the equation for the logit transformation of this logistic regression model. What characteristic of the outcome variable, MYOPIA, leads us to consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between MYOPIA and SPHEQ?

MYOPIA: Y (Y=0 if the subject has myopia; Y=1 if the subject does not have myopia)
SPHEQ: X

The logistic regression model:

$$\pi(x) = E(y \mid x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

The logit transformation of this logistic regression model:

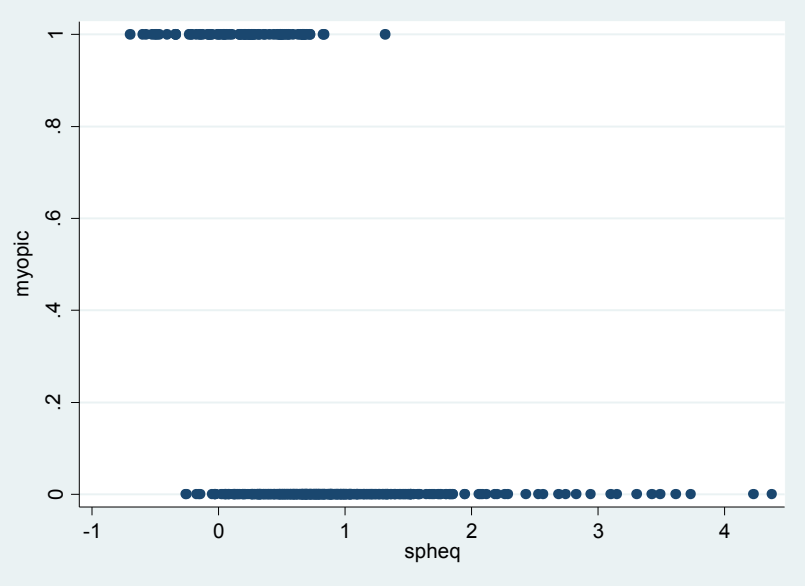
$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

MYOPIA is a binary variable, hence we consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between MYOPIA and SPHEQ.

**Notice that the left hand side of the logit transformation equation is the 'log-odds'*

b. Form a scatterplot of MYOPIA vs. SPHEQ.

Type 'twoway (scatter myopic spheq)' in the command window of Stata.



c. Write down an expression for the likelihood and log likelihood for the logistic regression model in part (a) using the ungrouped, $n = 618$, data. Obtain expressions for the two likelihood equations.

Likelihood function:

$$\ell(\beta) = \prod_{i=1}^{n} \left[\pi(x_i)\right]^{y_i} \left[1 - \pi(1 - x_i)\right]^{1-y_i}$$

Log-likelihood function:

$$L(\beta) = \ln\left(\ell(\beta)\right) = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\}$$

Likelihood equations:

$$\sum_{i=1}^{n} \left[y_i - \pi(x_i) \right] = 0$$

$$\sum_{i=1}^{n} x_i \left[y_i - \pi(x_i) \right] = 0$$

d. Using Stata, obtain the maximum likelihood estimates of the parameters of the logistic regression model in part (a). These estimates should be based on the ungrouped, $n = 618$, data. Using these estimates, write down the equation for the fitted values, that is, the estimated logistic probabilities. Plot the equation for the fitted values on the axes used in the scatterplots in parts (b) and (c).

Type 'logit myopic spheq' in the command window of Stata.

```
. logit  myopic spheq

Iteration 0:   log likelihood = -240.03851
Iteration 1:   log likelihood = -185.66235
Iteration 2:   log likelihood = -168.99543
Iteration 3:   log likelihood = -168.67355
Iteration 4:   log likelihood = -168.67244
Iteration 5:   log likelihood = -168.67244

Logistic regression                               Number of obs   =        618
                                                  LR chi2(1)      =     142.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -168.67244                       Pseudo R2       =     0.2973

------------------------------------------------------------------------------
      myopic |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       spheq |  -3.833098   .4183899    -9.16   0.000    -4.653127   -3.013068
       _cons |   .0539731    .206752     0.26   0.794    -.3512533    .4591996
------------------------------------------------------------------------------
```

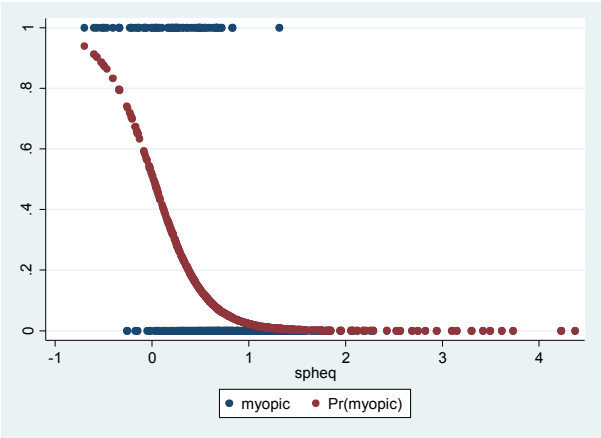
Below is the equation of the estimated logistic probabilities. Note that you need to exponentiate the logit estimates in order to obtain the logistic estimates.

$$\hat{\pi}(x) = \frac{\exp(0.05 - 3.83x)}{1 + \exp(0.05 - 3.83x)}$$

Type 'predict p' in the command window of Stata.
Then type 'twoway (scatter myopic spheq) (scatter p spheq)' in the command window.

```
. predict p
(option pr assumed; Pr(myopic))

. twoway (scatter myopic spheq) (scatter p
spheq)
```



Homework Solutions
Applied Logistic Regression

WEEK 1

Exercise 2:

Use the ICU study (icu.dta)

The ICU data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. A number of publications have appeared which have focused on various facets of this problem.

a. Write down the equation for the logistic regression model of STA on AGE. Write down the equation for the logit transformation of this logistic regression model. What characteristic of the outcome variable, STA, leads us to consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between STA and AGE?

Logistic regression model:

$$\pi(AGE) = \frac{e^{\beta_0+\beta_1 AGE}}{1+e^{\beta_0+\beta_1 AGE}}$$

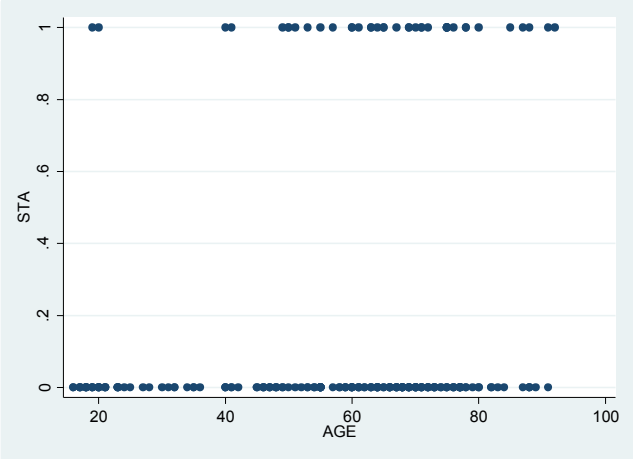
Logit transformation:

$$g(AGE) = \beta_0 + \beta_1(AGE)$$

We consider the logistic regression model, rather than the usual linear regression model to describe the relationship between STA and AGE because the outcome variable, STA, is dichotomous, taking on the values 0 and 1.

b. Form a scatterplot of STA versus AGE.

Type 'twoway (scatter STA AGE)' in the Stata command window.



c. Write down an expression for the likelihood and log likelihood for the logistic regression model in part (a) using the ungrouped, $n = 200$, data. Obtain expressions for the two likelihood equations.

likelihood function:

$$\ell(\beta) = \prod_{i=1}^{n} \zeta(x_i) \qquad \text{where,} \quad \zeta(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

x = AGE and $y_i = \begin{cases} 0 & \text{if the patient lived} \\ 1 & \text{if the patient died} \end{cases}$

log likelihood function:

$$L(\beta) = \ln\left(\ell(\beta)\right) = \sum_{i=1}^{n}\left\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\right\}$$

likelihood equations:

$$\sum_{i=1}^{n}\left[y_i - \pi(x_i)\right] = 0$$

$$\sum_{i=1}^{n} x_i\left[y_i - \pi(x_i)\right] = 0$$

d. Using Stata, obtain the maximum likelihood estimates of the parameters of the logistic regression model in part (a). These estimates should be based on the ungrouped, $n = 200$, data. Using these estimates, write down the equation for the fitted values, that is, the estimated logistic probabilities. Plot the equation for the fitted values on the axes used in the scatterplots in part (b).

Type 'logit STA AGE' in the command window of Stata.

```
. logit STA AGE

Iteration 0:   log likelihood = -100.08048
Iteration 1:   log likelihood = -96.261839
Iteration 2:   log likelihood =  -96.15328
Iteration 3:   log likelihood =  -96.15319
Iteration 4:   log likelihood =  -96.15319

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       7.85
                                                  Prob > chi2     =     0.0051
Log likelihood =  -96.15319                       Pseudo R2       =     0.0392

------------------------------------------------------------------------------
         STA |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |   .0275426   .0105647     2.61   0.009     .0068362     .048249
       _cons |  -3.058513    .696122    -4.39   0.000    -4.422887   -1.694139
------------------------------------------------------------------------------
```

logistic regression model (fitted values):

$$\pi(AGE) = \frac{e^{-3.058513+0.0275426\times(AGE)}}{1+e^{-3.058513+0.0275426\times(AGE)}}$$

logit transformation:

$$g(AGE) = -3.058513 + 0.0275426 \times (AGE)$$

To generate the fitted values using the equation for the fitted values, type in the Stata command window: 'generate NUM=exp(-3.0585+0.027542*(AGE))

Then type 'generate DEN=1+NUM'

Then type 'generate PROB=NUM/DEN'

Then type 'list AGE NUM DEN PROB'

```
. generate NUM=exp(-3.0585+0.027542*(AGE))

. list AGE NUM

           AGE         NUM
  1.        16    .0729612
  2.        16    .0729612
  3.        17    .0749986
  4.        17    .0749986
  5.        17    .0749986
…etc

. generate DEN=1+NUM

. generate PROB=NUM/DEN

. list AGE NUM DEN PROB

           AGE         NUM         DEN        PROB
  1.        16    .0729612    1.072961    .0679998
  2.        16    .0729612    1.072961    .0679998
  3.        17    .0749986    1.074999    .0697662
  4.        17    .0749986    1.074999    .0697662
  5.        17    .0749986    1.074999    .0697662
…etc
```

Alternatively, the probabilities can be generated by Stata directly after running the logistic regression model: type 'quietly logit STA AGE' in the command window; then type 'predict pihat'.

```
. quietly logit STA
AGE

. predict pihat
(option p assumed;
Pr(STA))
```

To plot the fitted curve of the logistic regression model on the axes used in the scatterplot in Exercise 2(b), type 'scatter STA pihat AGE, xscale(range10 100))

```
. scatter STA pihat AGE,
xscale(range(10 100))
```

e. Using the homework forum for this week, summarize (describe in words) the results presented in the plot obtained from parts (b) and (d).

The plot of STA vs. AGE (indicated in the scatterplot in 1(b)) demonstrates the dichotomous nature of the STA variable, which takes on the value zero if a patient is discharged alive or the value one if the patient died prior to discharge. The plot suggests that older people are more likely to die in the ICU, although overall, people are more likely to live than to die.

The plot of the estimated logistic probabilities vs. AGE (in Exercise 1(d)) indicates that the probability of dying does increase with increasing age. The rate of increase in the probabilities seems to increase with increasing age.