Solution to HW5

Problem 4.1

(a) The fitted model is

$$\log\{\widehat{\pi}/(1-\widehat{\pi})\} = -3.7771 + 0.1449 \times LI \Rightarrow \widehat{\pi} = \frac{e^{-3.7771+0.1449\times LI}}{1+e^{-3.7771+0.1449\times LI}}.$$

When $LI = 8$, $\widehat{\pi} = e^{-2.6179}/(1+e^{-2.6179}) = 0.068$.

(b) $\widehat{\pi} = 0.5 \Leftrightarrow \log\{\widehat{\pi}/(1-\widehat{\pi})\} = 0$, so $LI = 3.7771/0.1449 = 26.0$.

(c) The change rate at $LI = 8$ is

$$\beta\widehat{\pi}(LI = 8)\{1-\widehat{\pi}(LI = 8)\} = 0.1449 \times 0.068 \times (1-0.068) = 0.009.$$

The change rate at $LI = 26$ is

$$\beta\widehat{\pi}(LI = 26)\{1-\widehat{\pi}(LI = 26)\} = 0.1449 \times 0.5 \times (1-0.5) = 0.036.$$

(d) At $LI = 14$, the $\widehat{\pi}$ is

$$\widehat{\pi}(LI = 14) = \frac{e^{-3.7771+0.1449\times 14}}{1+e^{-3.7771+0.1449\times 14}} = 0.15.$$

At $LI = 28$, the $\widehat{\pi}$ is

$$\widehat{\pi}(LI = 28) = \frac{e^{-3.7771+0.1449\times 28}}{1+e^{-3.7771+0.1449\times 28}} = 0.57.$$

Therefore, the change of $\widehat{\pi}$ is 0.57-0.15=0.42 when LI changes from 14 to 28.

(e) When LI increases by 1, the odds multiplies by $e^{0.1449} = 1.16$ or increases by 16%.

Problem 4.2

(a) The Wald test statistic is

$$\chi^2 = \left(\frac{0.1449}{0.0593}\right)^2 = 5.97, \quad \text{P-value} = P(\chi^2_1 \geq 5.97) = 0.015.$$

Therefore, we conclude that there is a significant LI effect on the remission probability at 0.05 significant level.

(b) A 95% Wald CI for the odds-ratio e^{β} associated with one unit increase of LI:

$$[e^{0.1449-1.96\times 0.0593}, e^{0.1449+1.96\times 0.0593}] = [1.03, 1.30].$$

Interpretation: We are 95% confident that the odds-ratio associated with one unit increase of LI is at least 1.03 and at most 1.30.

(c) The LRT statistics for testing LI effect is $G^2 = 8.30$ with $df = 1$. So the P-value $= P(\chi_1^2 \geq 8.30) = 0.004$. Therefore, the LRT indicates that there is a significant LI effect(at 0.05 level, or any level > 0.004).

(d) A 95% LR CI for the odds-ratio e^{β} associated with one unit increase of LI: $[e^{0.0425}, e^{0.2846}] = [1.04, 1.33]$.

Interpretation: We are 95% confident that the odds-ratio associated with one unit increase of LI is at least 1.04 and at most 1.33.

Problem 4.7

(a) We fit the logistic model with age as the predictor as in the following:

```
data kyph;
  input age @@;
  y=1;
  datalines;
  12 15 42 52 59 73 82 91 96 105 114 120 121 128 130 139 139 157
  ;

data nokyph;
  input age @@;
  y=0;
  datalines;
  1 1 2 8 11 18 22 31 37 61 72 81 97 112 118 127 131 140 151 159 177 206
  ;

data main; set nokyph kyph;
run;

proc logistic descending;
  model y = age;
run;
***************************************************************************
                    Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio         0.5469        1         0.4596
            Score                    0.5444        1         0.4606
            Wald                     0.5393        1         0.4627

                   Analysis of Maximum Likelihood Estimates

                                           Standard          Wald
      Parameter     DF     Estimate           Error    Chi-Square     Pr > ChiSq

      Intercept      1      -0.5727          0.6024        0.9038         0.3418
      age            1      0.00430         0.00585        0.5393         0.4627
```

The LRT, score and Wald tests all indicate that there is no significant age effect (P-value > 0.4).

(b) The box plots of age for patients for whom kyphosis was absent and present are:

```
proc univariate plot;
  var age;
  by y;
run;
```

```
************************************************************************
                                Schematic Plots

                      |
                  200 +            |
                      |            |
                      |            |
                      |            |           |
                  150 +            |           |
                      |            |           |
                      |         +-----+     +-----+
                      |         |     |     |     |
                  100 +         |     |     *-----*
                      |         |     |     |  +  |
                      |         *--+--*     |     |
                      |         |     |     +-----+
                   50 +         |     |        |
                      |         |     |        |
                      |         |     |        |
                      |         +-----+        |
                    0 +            |
                       ------------+-----------+------------
                 y                 0           1
```

We see from the data plot that age distribution for patients without kyphosis has greater variance than the age distribution for patients with kyphosis. Therefore, if the age distributions are normal for each group, then in the logistic regression using age to predict kyphosis, there should be a quadratic term for age.

(c) We fit $\text{logit}[\pi(age)] = \alpha + \beta_1 age + \beta_2 age^2$ as in the following:

```
proc logistic descending;
  model y = age age*age;
run;
```

```
******************************************************************
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.0463	0.9944	4.2348	0.0396
age	1	0.0600	0.0268	5.0259	0.0250
age*age	1	-0.00033	0.000156	4.3960	0.0360

The Wald test for $H_0 : \beta_2 = 0$ is $\chi^2 = 4.4$ with P-value=0.036, significant at level 0.05!

The fit indicates that the logit of kyphosis probability is a quadratic function of age. As age increases from 1 month to $0.06/(2 \times 0.00033) = 91$ months, the log-odds increases from -1.97 to 0.68, and the probability of having kyphosis increases from 0.122 to 0.66. As age increases from 99 months to 206, the odds decreases from 0.68 to -3.69, and the probability of having kyphosis decreases from 0.66 to 0.02.

Problem 4.12

(a) The fitted model is

$$\text{logit}\{P(\text{death penalty})\} = -3.5961 - 0.8678\text{def} + 2.4044\text{vic}.$$

Covariate values that make the linear combination $-3.5961 - 0.8678\text{def} + 2.4044\text{vic}$ have the highest value will make the "yes" probability highest. Obviously, def should be 0 and vic should be 1. The corresponding "yes" probability is

$$P(\text{death penalty}) = \frac{e^{-3.5961+2.4044}}{1+e^{-3.5961+2.4044}} = 0.23.$$

(b) Interpretation of 2.4044, the coefficient of vic: given defendant's race, the odds a defendant receiving the death penalty when the victim is white is $e^{2.4044} = 11$ times the odds a defendant receiving the death penalty when the victim is black.

(c) A 95% LR CI: $[e^{1.3068}, e^{3.7175}] = [3.69, 41.16]$. We are 95% confident that the odds ratio of the death penalty between a white victim and a black victim is at least 3.69 and at most 41.16.

(d) The LR test for the effect of victim's race by controlling defendant's race: $G^2 = 20.35$ with $df = 1$ and P-value < 0.0001, which is significant at level 0.0001.

Interpretation: Given defendant's race, the odds ratio of a defendant's death penalty verdict between a white victim and a black victim is significantly different from 1.

1. **Problem 4.100**

(a) The sample odds-ratios of lung cancer between a smoking level to its previous level are: 3.72, 2.01, 1.28, 1.72, 1.66.

(b) Under the assumption that the logit model

$$\text{logit}P(\text{Lung cancer}|score)\} = \alpha + \beta score$$

is true in population, the same model is also true for the observed data from a case-control study with a difference intercept and the same β.

```
data lcancer;
  input score number $ y y0;
  n = y+y0;
  datalines;
  0  None    7          61
  1  <5      55  129
  2  5-14    489  570
  3  15-24   475  431
  4  25-49   293  154
  5  50+     38  12
  ;

proc logistic;
  model y/n = score;
run;
*************************************************************************
                    Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio       132.5343        1         <.0001
            Score                  129.2327        1         <.0001
            Wald                   123.0047        1         <.0001

                   Analysis of Maximum Likelihood Estimates

                                          Standard          Wald
      Parameter    DF     Estimate       Error     Chi-Square    Pr > ChiSq

      Intercept     1      -1.2151       0.1165      108.8645        <.0001
      score         1       0.4671       0.0421      123.0047        <.0001
```

The estimate of β is $\widehat{\beta} = 0.4671$ and a 95% Wald CI is $[0.4671 - 1.96 \times 0.0421, 0.4671 + 1.96 \times 0.0421] = [0.385, 0.550]$.

(c) The estimated odds-ratio associated with one unit increase of the smoking level is $e^{0.4671} = 1.60$. It is in the middle of the sample odds-ratios we got in (a). It is closer to the sample odds-ratios when the smoking level is higher, but less close to the sample odds-ratios when the smoking level is lower, probably due toe smaller sample sizes.