Homework Solutions
Applied Logistic Regression

WEEK 6

Exercise 1

From the ICU data, use as the outcome variable vital status (STA) and CPR prior to ICU admission (CPR) as a covariate.

(a) Demonstrate that the value of the log-odds ratio obtained from the cross-classification of STA by CPR is identical to the estimated slope coefficient from the logistic regression of STA on CPR. Verify that the estimated standard error of the estimated slope coefficient for CPR obtained from the logistic regression package is identical to the square root of the sum of the inverse of the cell frequencies from the cross-classification of STA by CPR. Use either set of computations to obtain 95% CI for the odds ratio. What aspect concerning the coding of the variable CPR makes the calculations for the two methods equivalent?

Type “tab CPR STA” in the command window to get the output shown below. This table is then used to obtain the odds ratio in the usual manner. Compute the log odds ratio so that it is easy to compare when we get the coefficient.

```
. tab CPR STA

           | STA
       CPR |         0           1 |
Total
-----------+----------------------+---------
-
         0 |       154          33 |
187
         1 |         6           7 |
13
-----------+----------------------+---------
-
     Total |       160          40 |
200
```

$$OR = \frac{a*d}{b*c} = \frac{(154)*(7)}{(33)*(6)} = 5.44$$

$$\ln(OR) = 1.6946$$

Type "logit STA CPR" in the command window to fit the logistic regression model. The output so obtained is shown below.

```
. logit STA CPR

Iteration 0:  Log Likelihood =-100.08048
Iteration 1:  Log Likelihood =-96.634239
Iteration 2:  Log Likelihood =-96.117589
Iteration 3:  Log Likelihood =-96.114275

Logit Estimates                                        Number of obs =    200
                                                       chi2(1)       =   7.93
                                                       Prob > chi2   = 0.0049
Log Likelihood = -96.114275                            Pseudo R2     = 0.0396

------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     CPR |   1.694596   .5884886      2.880   0.004       .5411793    2.848012
   _cons |  -1.540445   .1918242     -8.031   0.000      -1.916414   -1.164476
------------------------------------------------------------------------------
```

The log-odds ratio from the 2x2 contingency table is identical to the estimated slope coefficient from the logistic regression of STA on CPR.

Verify that the estimated standard error of the estimated slope coefficient for CPR obtained from the logistic regression package is identical to the square root of the sum of the inverse of the cell frequencies from the cross-classification of STA by CPR.

From the output above, it is clear that the estimated standard error of the estimated slope coefficient from the logistic regression is:

$$\widehat{SE}\left(\hat{\beta}_1\right) = 0.588$$

This is equivalent to the square root of the sum of the inverse of the cell frequencies from the 2x2 contingency table:

$$\widehat{SE}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{154} + \frac{1}{33} + \frac{1}{6} + \frac{1}{7}} = 0.588$$

Use either set of computations to obtain 95% CI for the odds ratio.

$$95\%\text{CI} = \exp[\hat{\beta}_1 \pm z_{1-\alpha/2} * S\hat{E}\left(\hat{\beta}_1\right)]$$

$$= \exp[1.6946 \pm 1.96 * 0.588]$$

$$1.718 \le OR \le 17.237$$

This is equivalent to the 95% CI for the estimated odds ratio that can be obtained from the logistic procedure in STATA:

Type "logistic STA CPR" in the command window to obtain the output shown below. This command directly generates the odds ratio and also gives its standard error and confidence interval.

```
. logistic STA CPR

Logit Estimates                                   Number of obs =    200
                                                  chi2(1)       =   7.93
                                                  Prob > chi2   = 0.0049
Log Likelihood = -96.114275                       Pseudo R2     = 0.0396

------------------------------------------------------------------------------
     STA | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     CPR |   5.444444   3.203994      2.880   0.004       1.718032    17.25345
------------------------------------------------------------------------------
```

What aspect concerning the coding of the variable CPR makes the calculations for the two methods equivalent?

The calculations from the two methods are equivalent because the variable CPR is dichotomous and has been coded 0,1.

(b) For purposes of illustration, use a data transformation statement to recode, for this problem only, the variable CPR as follows: 4 = no and 2 = yes. Perform the logistic regression of STA on CPR (recoded). Demonstrate how the calculation of the logit difference of CPR = yes versus CPR = no is equivalent to the value of the log-odds ratio obtained in exercise 1-a. Use the results from the logistic regression to obtain the 95% CI for the odds ratio and verify that they are the same limits as obtained in Exercise 1-a.

To recode CPR use the following command. (cprnew is the recoded CPR and we will be using cprnew for our analysis)

```
. gen cprnew=2
. replace cprnew=4 if CPR==0
(187 real changes made)
```

Perform the logistic regression of STA on CPR (recoded).

Type "logit STA cprnew" in the command window to obtain the output below.

```
. logit STA cprnew

Iteration 0:  Log Likelihood =-100.08048
Iteration 1:  Log Likelihood =-96.634239
Iteration 2:  Log Likelihood =-96.117589
Iteration 3:  Log Likelihood =-96.114275

Logit Estimates                                   Number of obs =    200
                                                  chi2(1)       =   7.93
                                                  Prob > chi2   = 0.0049
Log Likelihood = -96.114275                       Pseudo R2     = 0.0396

------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  cprnew |  -.8472979   .2942443     -2.880   0.004      -1.424006   -.2705896
   _cons |   1.848746   1.129108      1.637   0.102      -.3642654    4.061758
------------------------------------------------------------------------------
```

Demonstrate how the calculation of the logit difference of CPR=yes vs. CPR=no is equivalent to the value of the log-odds ratio obtained in Exercise 1-a.

$$
\begin{aligned}
\ln\left[\widehat{\mathrm{OR}}(a,b)\right] &= \hat{g}(x=a)-\hat{g}(x=b) \\
&= \left(\hat{\beta}_0+\hat{\beta}_1 * a\right)-\left(\hat{\beta}_0+\hat{\beta}_1 * b\right) \\
&= \hat{\beta}_1 * (a-b) \\
&= -0.8472979 * (2-4) \\
&= 1.6946
\end{aligned}
$$

This is equivalent to the value of the log-odds ratio obtained in exercise 1-a.

Use the results from the logistic regression to obtain the 95% CI for the odds ratio and verify that they are the same limits as obtained in Exercise 1-a.

$$95\%\mathrm{CI} = \exp[2\hat{\beta}_1 \pm z_{1-\alpha/2} * 2\widehat{SE}(\hat{\beta}_1)]$$

Under this coding scheme, the independent variable takes on a lower value (2) when the covariate is present than when the covariate is absent (4). In order to preserve the comparison that was made in Exercise 1-a, the sign of the coefficient is changed from negative to positive. This will lead to the appropriate confidence interval for the odds ratio.

$$
\begin{aligned}
95\%\mathrm{CI} &= \exp[2*(0.8472979) \pm 1.96 * 2 * (0.2942443)] \\
&= \exp[1.6946 \pm 1.96 * 0.588] \\
&1.718 \le OR \le 17.237
\end{aligned}
$$

This is the same confidence interval for the odds ratio that was obtained in Exercise 1-a.

(c) Consider the ICU data and use as the outcome variable vital status (STA) and race (RACE) as a covariate. Prepare a table showing the coding of the two design variables for RACE using the value RACE = 1, white, as the reference group. Show that the estimated log-odds ratios obtained from the cross-classification of STA by RACE, using RACE = 1 as the reference group, are identical to estimated slope coefficients for the two design variables from the logistic regression of STA on RACE. Verify that the estimated standard errors of the estimated slope coefficients for the two design variables for RACE are identical to the square root of the sum of the inverse of the cell frequencies from the cross-classification of STA by RACE used to calculate the odds ratio. Use either set of computations to compute the 95% CI for the odds ratios.

The dummy coding for the race

RACE	Label	RACE_2	RACE_3
1	White	0	0
2	Black	1	0
3	Other	0	1

Show that the estimated log-odds ratios obtained from the cross-classification of STA by RACE, using RACE=1 as the reference group, are identical to estimated slope coefficients for the two design variables from the logistic regression of STA on RACE.

Type "tab RACE STA" in the command window to obtain a 2X2 table.

```
. tab RACE STA

           | STA
      RACE |         0          1 |
Total
-----------+----------------------+----
-
         1 |       138         37 |
175
         2 |        14          1 |
15
         3 |         8          2 |
10
-----------+----------------------+----
---
     Total |       160         40 |
200
```

$$OR(\text{Race}=2 \text{ vs. Race}=1) = \frac{a*d}{b*c} = \frac{(138)*(1)}{(37)*(14)} = 0.266$$

$$\ln(OR(2 \text{ vs. } 1)) = -1.3227$$

$$OR(\text{Race}=3 \text{ vs. Race}=1) = \frac{a*d}{b*c} = \frac{(138)*(2)}{(37)*(8)} = 0.9324$$

$$\ln(OR(3 \text{ vs. } 1)) = -0.069958$$

Type "xi: logit STA i.RACE" in the command window to obtain the logistic regression output below. This syntax is used whenever you have a dummy coded variable. Race_1 will be considered as the referent cell.

```
. xi: logit STA i.RACE
i.RACE                IRACE_1-3     (naturally coded; IRACE_1 omitted)

Iteration 0:   log likelihood = -100.08048
Iteration 1:   log likelihood = -99.043397
Iteration 2:   log likelihood = -98.952555
Iteration 3:   log likelihood =  -98.95055
Iteration 4:   log likelihood = -98.950549

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(2)      =       2.26
                                                  Prob > chi2     =     0.3231
Log likelihood = -98.950549                       Pseudo R2       =     0.0113

------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 IRACE_2 |  -1.322722   1.051523     -1.258   0.208      -3.383669    .7382257
 IRACE_3 |  -.0699586   .8119565     -0.086   0.931      -1.661364    1.521447
   _cons |  -1.316336   .1851308     -7.110   0.000      -1.679185   -.9534861
```

The estimated log-odds ratios obtained from the cross-classification of STA by RACE, using RACE=1 as the reference group, are identical to estimated slope coefficients for the two design variables from the logistic regression of STA on RACE.

Verify that the estimated standard errors of the estimated slope coefficients for the two design variables for RACE are identical to the square root of the sum of the inverse of the cell frequencies from the cross-classification of STA by RACE used to calculate the odds ratio.

For RACE 2 vs. RACE 1

$$\widehat{SE}\left(\hat{\beta}_1\right) = 1.051523 \qquad \text{from the estimated logit}$$

$$\widehat{SE}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{138} + \frac{1}{1} + \frac{1}{37} + \frac{1}{14}} = 1.051523$$

For RACE 3 vs. RACE 1

$$\widehat{SE}\left(\hat{\beta}_1\right) = 0.8119565 \qquad \text{from the estimated logit}$$

$$\widehat{SE}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{138} + \frac{1}{2} + \frac{1}{37} + \frac{1}{8}} = 0.8119565$$

Use either set of computations to compute the 95% CI for the odds ratios.

For RACE 2 vs. RACE 1

$$95\%\text{CI} = \exp[\hat{\beta}_1 \pm z_{1-\alpha/2} * \widehat{SE}\left(\hat{\beta}_1\right)]$$

$$= \exp[-1.3227 \pm 1.96 * \left(1.051523\right)]$$

$$0.03392 \le OR \le 2.0923$$

For RACE 3 vs. RACE 1

$$95\%\text{CI} = \exp[\hat{\beta}_1 \pm z_{1-\alpha/2} * \widehat{SE}\left(\hat{\beta}_1\right)]$$

$$= \exp[-0.0699586 \pm 1.96 * \left(0.811956\right)]$$

$$0.18987 \le OR \le 4.578979$$

These are equivalent to the 95% confidence intervals for the estimated odds ratios that can be obtained from the logistic procedure in STATA:

```
. xi: logistic STA i.RACE
i.RACE              IRACE_1-3    (naturally coded; IRACE_1 omitted)

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(2)      =       2.26
                                                  Prob > chi2     =     0.3231
Log likelihood = -98.950549                       Pseudo R2       =     0.0113

------------------------------------------------------------------------------
     STA | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 IRACE_2 |   .2664093   .2801355     -1.258   0.208        .0339228    2.09222
 IRACE_3 |   .9324324   .7570946     -0.086   0.931        .1898798   4.578846
------------------------------------------------------------------------------
```