Homework
Applied Logistic Regression

WEEK 3

Problem 3:

Use the ICU study (icu.dta)

Use the ICU data and consider the multiple logistic regression model of vital status, STA, on age (AGE), cancer part of the present problem (CAN), CPR prior to ICU admission (CPR), infection probable at ICU admission (INF), and race (RACE).

a. **The variable RACE is coded at three levels. Prepare a table showing the coding of the two design variables necessary for including this variable in a logistic regression model.**

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

b. **Write down the equation for the logistic regression model of STA on AGE, CAN, CPR, INF, and RACE. Write down the equation for the logit transformation of this logistic regression model. How many parameters does this model contain?**

The logistic regression model:

$$\pi\left(\mathbf{x}\right)=\frac{e^{\beta_0+\beta_1*(AGE)+\beta_2*\left(CAN\right)+\beta_3*\left(CPR\right)+\beta_4*\left(INF\right)+\beta_5*\left(RACE\_2\right)+\beta_6*\left(RACE\_3\right)}}{1+e^{\beta_0+\beta_1*(AGE)+\beta_2*\left(CAN\right)+\beta_3*\left(CPR\right)+\beta_4*\left(INF\right)+\beta_5*\left(RACE\_2\right)+\beta_6*\left(RACE\_3\right)}}$$

where $\mathbf{X}$ = vector of covariates

The logit transformation is:

$$g\left(\mathbf{x}\right)=\beta_0+\beta_1*(AGE)+\beta_2*\left(CAN\right)+\beta_3*\left(CPR\right)+\beta_4*\left(INF\right)+\beta_5*\left(RACE\_2\right)+\beta_6*\left(RACE\_3\right)$$

This model contains 7 parameters.

c. Write down an expression for the likelihood and log likelihood for the logistic regression model in part (b). How many likelihood equations are there? Write down an expression for a typical likelihood equation for this problem.

Likelihood function:

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

x = set of covariates and

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

log likelihood function:

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

There will be p+1 or 7 likelihood equations for this problem.

Likelihood equations that result may be expressed as follows:

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

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

d. Using a logistic regression package, obtain the maximum likelihood estimates of the parameters of the logistic regression model in part (b). Using these estimates write down the equation for the fitted values, that is, the estimated logistic probabilities.

The following code (xi: logit STA AGE CAN CPR INF i.RACE) will tell Stata to automatically create dummy variables for RACE: IRACE_2 and IRACE_3.

```
. xi: logit STA AGE CAN CPR INF i.RACE

i.RACE            _IRACE_1-3          (naturally coded; _IRACE_1 omitted)

Iteration 0:   log likelihood = -100.08048
Iteration 1:   log likelihood = -90.619912
Iteration 2:   log likelihood = -89.663593
Iteration 3:   log likelihood = -89.650384
Iteration 4:   log likelihood = -89.650364     -2(-89.650364)=179.3007 : Deviance

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(6)      =      20.86
                                                  Prob > chi2     =     0.0019
Log likelihood = -89.650364                       Pseudo R2       =     0.1042

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

```
         STA |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |   .0271207   .0115879     2.34   0.019     .0044089    .0498325
         CAN |   .2445106    .616815     0.40   0.692    -.9644246    1.453446
         CPR |   1.646497   .6234135     2.64   0.008      .424629    2.868365
         INF |   .6806676   .3804176     1.79   0.074    -.0649372    1.426272
    _IRACE_2 |  -.9570777   1.084467    -0.88   0.377    -3.082593    1.168438
    _IRACE_3 |   .2597493   .8712682     0.30   0.766    -1.447905    1.967404
       _cons |   -3.51152   .8144295    -4.31   0.000    -5.107772   -1.915267

. estimates store A
```

The logit transformation:

$$g\left(x\right)=-3.512+0.027*(AGE)+0.245*(CAN)+1.646(CPR)+0.681\left(INF\right)-0.957\left(RACE\_2\right)+0.260\left(RACE\_3\right)$$

The logistic regression model:

$$\pi\left(x\right)=\frac{e^{-3.512+0.027*(AGE)+0.245*(CAN)+1.646(CPR)+0.681(INF)-0.957*(RACE\_2)+0.260(RACE\_3)}}{1+e^{-3.512+0.027*(AGE)+0.245*(CAN)+1.646(CPR)+0.681(INF)-0.957*(RACE\_2)+0.260(RACE\_3)}}$$

e. Using the results of the output from the logistic regression package used in part (d), assess the significance of the slope coefficients for the variables in the model using the likelihood ratio test. What assumptions are needed for the p-values computed for this test to be valid? What is the value of the deviance for the fitted model?

Deviance:

$$D=-2\ln\left[\frac{\left(\text{likelihood of the current model}\right)}{\left(\text{likelihood of the saturated model}\right)}\right]$$

$D=-2\left(-89.650364\right)$

$D=179.300728$

Likelihood Ratio Test:

$$H_0:\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=0$$

H_A: At least one coefficient is not equal to 0

$$G=D\left(\text{model without variable}\right)-D\left(\text{model with variable}\right)$$

$G=200.16-179.300728$

$G = 20.86$ $G \sim \chi^2(6)$ $p = 0.00194$

$\therefore$ reject H$_0$, it is not consistent with the data that all β=0; we conclude that together, AGE, CAN, CPR, INF, and RACE are significant predictors of STA. Assumption: the statistic G will follow a χ^2 distribution with 6 degrees of freedom under the null hypothesis.

The value of the deviance for the fitted model is: $D = 179.30$.

f. Use the Wald statistics to obtain an approximation to the significance of the individual slope coefficients for the variables in the model. Fit a reduced model that eliminates those variables with nonsignificant Wald statistics. Assess the joint (conditional) significance of the variables excluded from the model. Present the results of fitting the reduced model in a table.

Using a critical value of $p < 0.05$, based on the computer output from Exercise 1(d), one can conclude that the variables AGE, CPR and possibly INF are significant while CAN, RACE_2 and RACE_3 are not significant. A reduced model was fit containing only those variables thought to be significant:

```
. logit STA AGE CPR INF

Iteration 0:  Log Likelihood =-100.08048
Iteration 1:  Log Likelihood =-91.107974
Iteration 2:  Log Likelihood =-90.262687
Iteration 3:  Log Likelihood =-90.256715
Iteration 4:  Log Likelihood =-90.256714

Logit Estimates                                        Number of obs =    200
                                                       chi2(3)       =  19.65
                                                       Prob > chi2   = 0.0002
Log Likelihood = -90.256714                            Pseudo R2     = 0.0982

------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------
     age |    .027922   .0113598      2.458   0.014       .0056573    .0501867
     CPR |   1.630662   .6155313      2.649   0.008       .4242431    2.837081
     inf |   .6970764      .3775      1.847   0.065      -.0428101    1.436963
   _cons |  -3.576045   .7730606     -4.626   0.000      -5.091216   -2.060874

. lrtest A .
Logit:  likelihood-ratio test                        chi2(3)     =       1.21
                                                     Prob > chi2 =     0.7500
```

The likelihood ratio test comparing the above model with the full model will have a distribution that is chi-square with 3 degrees of freedom under the hypothesis that the coefficients for the variables that were excluded are equal to zero.

Likelihood Ratio Test:

$$H_0: \text{Coefficients for eliminated variables all equal 0}$$

$$H_A: \text{At least one coefficient is not equal to 0}$$

$$G = D(\text{model without variables}) - D(\text{model with variables})$$

$$G = 180.513428 - 179.300728$$

$G = 1.21$ $G \sim \chi^2(3)$ $p = 0.75061$

∴ do not reject H_0, it is consistent with the data that the coefficients for the eliminated variables are all equal to zero; we conclude that the reduced model containing only AGE, CPR and INF is as good as the full model. Statistically speaking, there is no advantage to including CAN or RACE in the model.

Based on a strict definition of significance at $p<0.05$, one would also exclude the variable INF from the model. The reduced model contains AGE and CPR only.

```
. logit STA AGE CPR

Iteration 0:  Log Likelihood =-100.08048
Iteration 1:  Log Likelihood =-92.714062
Iteration 2:  Log Likelihood = -91.98047
Iteration 3:  Log Likelihood = -91.97634
Iteration 4:  Log Likelihood = -91.97634

Logit Estimates                                        Number of obs =     200
                                                       chi2(2)       =   16.21
                                                       Prob > chi2   =  0.0003
Log Likelihood =  -91.97634                            Pseudo R2     =  0.0810

------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .0296074   .0111489      2.656   0.008       .0077559    .0514589
     CPR |   1.784092   .6072971      2.938   0.003       .5938116    2.974373
   _cons |  -3.351956   .7454995     -4.496   0.000      -4.813108   -1.890803
------------------------------------------------------------------------------
```

The likelihood ratio test comparing the above model with the model that included INF will have a distribution that is chi-square with 1 degree of freedom under the hypothesis that the coefficient for INF is equal to zero.

Likelihood Ratio Test:

$$H_0 : \text{Coefficient for INF equals } 0$$
$$H_A : \text{Coefficient for INF is not equal to } 0$$
$$G = D(\text{model without variable}) - D(\text{model with variable})$$
$$G = 183.95 - 180.51$$
$$G = 3.44 \qquad G \sim \chi^2(1) \qquad p = 0.06364$$

∴ do not reject H_0, it is consistent with the data that the coefficient for the eliminated variable, INF, is equal to zero; we conclude that the reduced model containing only AGE and CPR is as good as the full model. Statistically speaking, there is no advantage to including INF in the model. However, the significance of INF is borderline. The decision to include INF should be made after considering clinical reasons for its inclusion in the model.

g. Using the results from part (f), compute 95 percent confidence intervals for all coefficients in the model. Write a sentence interpreting the confidence intervals for the non-constant covariates.

Ninety-five percent confidence intervals for the coefficients for AGE and CPR can be computed using equations (1.15) and (1.16) from the text. The intervals shown below are calculated from the STATA output for the logistic regression model shown in problem 1(f). The confidence intervals for the coefficients can also be seen in the 6th and 7th columns presented in this STATA output.

Endpoints of a 100(1-α)% confidence interval for AGE:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}\left(\hat{\beta}_1\right)$$

$$0.0296 \pm 1.96(0.0111)$$

$$(0.0078,\ 0.0515)$$

Endpoints of a 100(1-α)% confidence interval for CPR:

$$\hat{\beta}_2 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_2)$$

$$1.7841 \pm 1.96(0.6073)$$

$$(0.5938,\ 2.9744)$$

Endpoints of a 100(1-α)% confidence interval for constant:

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_0)$$

$$-3.3520 \pm 1.96(0.7455)$$

$$(-4.8131,\ -1.8908)$$

The 95% confidence interval for AGE suggests that the change in the log odds of dying in the ICU (STA=1) per one year increase in AGE is 0.0296 when the value of CPR is constant and that the change could be as little as 0.0078 or as much as 0.0515 with 95% confidence.

The 95% confidence interval for CPR suggests that the change in the log odds of dying in the ICU (STA=1) for persons who had CPR prior to admission compared with those who had not is 1.7841 when the value of AGE is constant and that the change could be as little as 0.5938 or as much as 2.9744 with 95% confidence.