Homework Solutions
Applied Logistic Regression

WEEK 6

Exercise 1

(d) Create design variables for RACE using the method typically employed in ANOVA. Perform the logistic regression of STA on RACE. Show by calculation that the estimated logit differences of RACE = 2 versus RACE = 1 and RACE = 3 versus RACE = 1 are equivalent to the values of the log-odds ratio obtained in problem 1(c). Use the results of the logistic regression to obtain the 95% CI for the odds ratios and verify that they are the same limits as obtained in Exercise 1(c). Note that the estimated covariance matrix for the estimated coefficients is needed to obtain the estimated variances of the logit differences.

Use the following codes to create ANOVA like design variables for RACE

```
. gen rdvm_1=-1

. replace rdvm_1=1 if RACE==2
(15 real changes made)

. replace rdvm_1=0 if RACE==3
(10 real changes made)

. gen rdvm_2=-1

. replace rdvm_2=0 if RACE==2
(15 real changes made)

. replace rdvm_2=1 if RACE==3
(10 real changes made)
```

RACE	Label	RDVM_1	RDVM_2
1	White	-1	-1
2	Black	1	0
3	Other	0	1

Perform the logistic regression of STA on RACE. (use the new design variables of RACE)

Type "logit STA rdvm_1 rdvm_2" in the command window to obtain the logistic regression output.

```
. logit STA rdvm_1 rdvm_2

Logit Estimates                                       Number of obs =    200
                                                      chi2(2)       =   2.26
                                                      Prob > chi2   = 0.3231
Log Likelihood = -98.950549                           Pseudo R2     = 0.011
------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  rdvm_1 |  -.8584948   .7412439     -1.158   0.247      -2.311306    .5943165
  rdvm_2 |   .3942681   .6329561      0.623   0.533       -.846303    1.634839
   _cons |  -1.780562   .4385203     -4.060   0.000      -2.640047   -.9210784
------------------------------------------------------------------------------
```

Show by calculation that the estimated logit differences of RACE=2 vs. RACE=1 and RACE=3 vs. RACE=1 are equivalent to the values of the log-odds ratio obtained in problem 2.1.

For RACE 2 vs. RACE 1

$$\ln\left[\widehat{OR}\left(\text{black,white}\right)\right] = \hat{g}\left(\text{black}\right) - \hat{g}\left(\text{white}\right)$$

$$= \hat{\beta}_0 + \hat{\beta}_{11} * (D_1 = 1) + \hat{\beta}_{12} * \left(D_2 = 0\right)$$

$$-\left[\hat{\beta}_0 + \hat{\beta}_{11} * (D_1 = -1) + \hat{\beta}_{12} * \left(D_2 = -1\right)\right]$$

$$= 2\hat{\beta}_{11} + \hat{\beta}_{12}$$

$$= 2 * (-0.8584948) + 0.3942681$$

$$\ln\left[\widehat{OR}\left(\text{black,white}\right)\right] = -1.322722$$

For RACE 3 vs. RACE 1

$$\ln\left[\widehat{OR}\left(\text{other,white}\right)\right] = \hat{g}\left(\text{other}\right) - \hat{g}\left(\text{white}\right)$$

$$= \hat{\beta}_0 + \hat{\beta}_{11} * (D_1 = 0) + \hat{\beta}_{12} * \left(D_2 = 1\right)$$

$$-\left[\hat{\beta}_0 + \hat{\beta}_{11} * (D_1 = -1) + \hat{\beta}_{12} * \left(D_2 = -1\right)\right]$$

$$= \hat{\beta}_{11} + 2\hat{\beta}_{12}$$

$$= (-0.8584948) + 2 * \left(0.3942681\right)$$

$$= -0.0699586$$

Use the results of the logistic regression to obtain 95% confidence intervals for the ORs and verify that they are the same limits as obtained in problem 1(c).

Type "vce" in the command window to obtain the variance-covariance matrix.

```
. vce

        |    rdvm_1    rdvm_2
_cons
--------+-------------------------
-
  rdvm_1|   .549443
  rdvm_2| -.373176    .400633
   _cons|   .164842    .016033
.1923
```

For RACE 2 vs. RACE 1

$$\widehat{Var}\left\{\ln\left[\widehat{OR}(\text{black, white})\right]\right\} = 4\widehat{var}\left(\hat{\beta}_{11}\right) + \widehat{var}\left(\hat{\beta}_{12}\right) + 4\widehat{\text{cov}}\left(\hat{\beta}_{11},\hat{\beta}_{12}\right)$$

$$= 4(0.549443) + (0.400633) + 4(-0.373176)$$

$$= 1.105701$$

$$\widehat{SE} = \sqrt{1.105701} = 1.0515$$

$$95\%\text{CI} = \exp[2\hat{\beta}_{11} + \hat{\beta}_{12} \pm z_{1-\alpha/2} * \widehat{SE}\left(2\hat{\beta}_{11} + \hat{\beta}_{12}\right)]$$

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

$$0.03392 \le OR \le 2.0923$$

For RACE 3 vs. RACE 1

$$\widehat{Var}\left\{\ln\left[\widehat{OR}(\text{other, white})\right]\right\} = \widehat{\text{var}}\left(\hat{\beta}_{11}\right) + 4\widehat{\text{var}}\left(\hat{\beta}_{12}\right) + 4\widehat{\text{cov}}\left(\hat{\beta}_{11},\hat{\beta}_{12}\right)$$

$$= (0.549443) + 4(0.400633) + 4(-0.373176)$$

$$= 0.659271$$

$$\widehat{SE} = \sqrt{0.659271} = 0.811955048$$

$$95\%\text{CI} = \exp[\hat{\beta}_{11} + 2\hat{\beta}_{12} \pm z_{1-\alpha/2} * \widehat{SE}\left(\hat{\beta}_{11} + 2\hat{\beta}_{12}\right)]$$

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

$$0.18987 \le OR \le 4.578979$$

These are the same confidence intervals as the ones obtained in Exercise 1-c.

(e) Consider the logistic regression of STA on CRN and AGE. Consider CRN to be the risk factor and show that AGE is a confounder of the association of CRN with STA. Addition of the interaction of AGE by CRN presents an interesting modeling dilemma. Examine the main effects only and interaction models graphically. Using the graphical results and any significance tests you feel are needed, select the best model (main effects or interaction) and justify your choice. Estimate relevant odds ratios. Repeat this analysis of confounding and interaction for a model that includes CPR as the risk factor and AGE as the potential confounding variable.

```
. logit STA CRN

Logit Estimates                                        Number of obs =
200
                                                       chi2(1)       =
5.42
                                                       Prob > chi2   =
0.0199
Log Likelihood = -97.368374                            Pseudo R2     =
0.0271

 ------------------------------------------------------------------------
                                                                  -------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf.
Interval]
---------+---------------------------------------------------------------
-------
     CRN |   1.219757   .5038556      2.421   0.015       .2322178
2.207296
   _cons |   -1.53821   .1948369     -7.895   0.000      -1.920084   -
1.156337
-------------------------------------------------------------------------
-------

. logit STA CRN AGE


Logit Estimates                                        Number of obs =
200
                                                       chi2(2)       =
11.56
                                                       Prob > chi2   =
0.0031
Log Likelihood = -94.302294                            Pseudo R2     =
0.0577

-------------------------------------------------------------------------
-------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf.
Interval]
---------+---------------------------------------------------------------
-------
     CRN |   1.019856   .5149228      1.981   0.048       .0106262
2.029087
     AGE |   .0249915   .0107232      2.331   0.020       .0039744
.0460085
   _cons |  -3.029875   .7000099     -4.328   0.000       -4.40187   -
1.657881

. predict xbtest, xb
```

There is a 16.4% decrease in the value for the coefficient for CRN when AGE is adjusted for in the model. A decrease of this magnitude indicates that AGE confounds the relationship between CRN and STA.

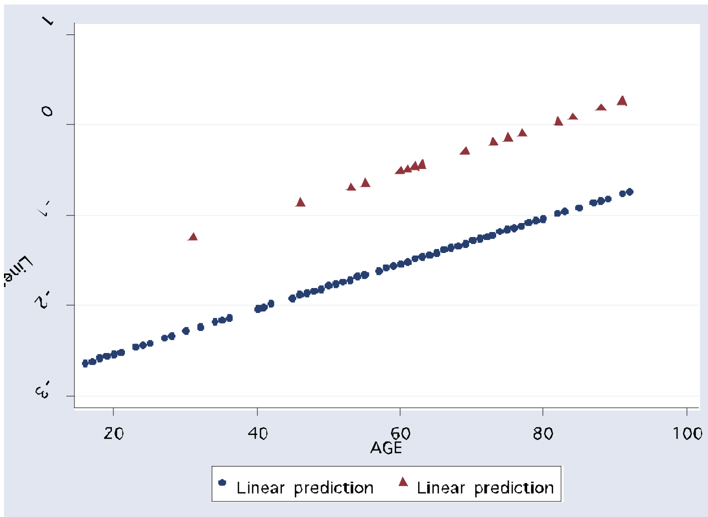
Addition of the interaction of AGE by CRN presents an interesting modeling dilemma.

```
. gen crnage=CRN*AGE
. logit STA CRN AGE crnage

Iteration 0:  Log Likelihood =-100.08048
Iteration 1:  Log Likelihood =-94.160869
Iteration 2:  Log Likelihood =-93.683315
Iteration 3:  Log Likelihood =-93.681076
Iteration 4:  Log Likelihood =-93.681076

Logit Estimates                                        Number of obs =    200
                                                       chi2(3)       =  12.80
                                                       Prob > chi2   = 0.0051
Log Likelihood = -93.681076                            Pseudo R2     = 0.0639

------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     CRN |   3.573101   2.322261      1.539   0.124      -.9784469    8.124649
     AGE |    .029242    .011725      2.494   0.013       .0062613    .0522226
  crnage |  -.0380925   .0340579     -1.118   0.263      -.1048447    .0286598
   _cons |  -3.297927   .7705345     -4.280   0.000      -4.808147   -1.787707


. predict xbtest1, xb
```

Examine the main effects only and interaction models graphically.

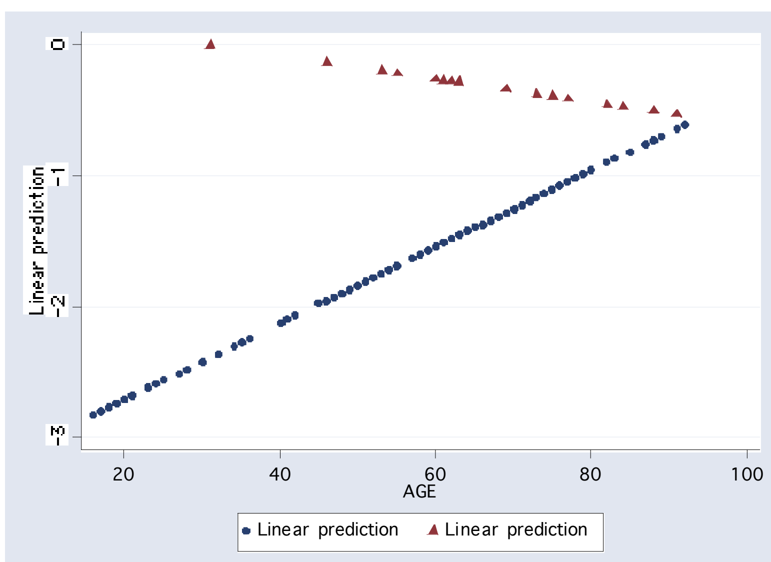
Type the following in the command window to obtain the scatter plot.

```
. scatter xbtest AGE if CRN==0, msymbol(circle) || scatter xbtest AGE if CRN==1,
 msymbol(triangle)
```



Scatterplot of the main effects model for STA on CRN AGE
logit vs. AGE by CRN
diamond for logit when CRN=0
triangle for logit when CRN

```
. scatter xbtest1 AGE if CRN==0, msymbol(circle) || scatter
xbtest1 AGE if  CRN==1, msymbol(triangle)
```



Scatterplot of the interaction model for STA on CRN AGE CRN*AGE
logit vs. AGE by CRN
diamond for logit when CRN=0
triangle logit when CRN=1

Using the graphical results and any significance tests you feel are needed, select the best model (main effects or interaction) and justify your choice. Estimate the relevant odds ratios.

Model	Constant	CRN	AGE	CRN*AGE	log-likelihood	G	p-value
1	-1.53821	1.219757	-		-97.34		
2	-3.029875	1.019856	0.0249915	-	-94.30	6.08	0.014
3	-3.297927	3.573101	0.029242	-0.038093	-93.68	1.24	0.265

Based on the impression gained from looking at the graph of the logits from the model containing no interaction term, as well as the Wald statistic for the interaction term (crnage) and the results of the likelihood ratio test, it does not appear justified to include the interaction term in the model. There is no effect modification by AGE.

The relevant odds ratio (adjusted for AGE) can be obtained by exponentiating the coefficient for CRN from the model that includes AGE as a covariate:

$$\widehat{OR} = \exp(1.019856) = 2.77$$

The 95% CI for this odds ratio can be obtained by exponentiating the endpoints of the 95% confidence interval for the coefficient for CRN from the model that includes AGE as a covariate:

$$\exp\left(0.0106262\right) \le OR \le \exp\left(2.029087\right)$$

$$1.01 \le OR \le 7.61$$

Repeat this analysis of confounding and interaction for a model which includes CPR as the risk factor and AGE as the potential confounding variable.

```
. logit 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 |      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
------------------------------------------------------------------------------

. logit STA CPR AGE

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]
---------+--------------------------------------------------------------------
     CPR |   1.784092   .6072971      2.938   0.003       .5938116    2.974373
     AGE |   .0296074   .0111489      2.656   0.008       .0077559    .0514589
```

```
    _cons |  -3.351956   .7454995     -4.496   0.000      -4.813108   -1.890803

. predict xbtest3, xb
```

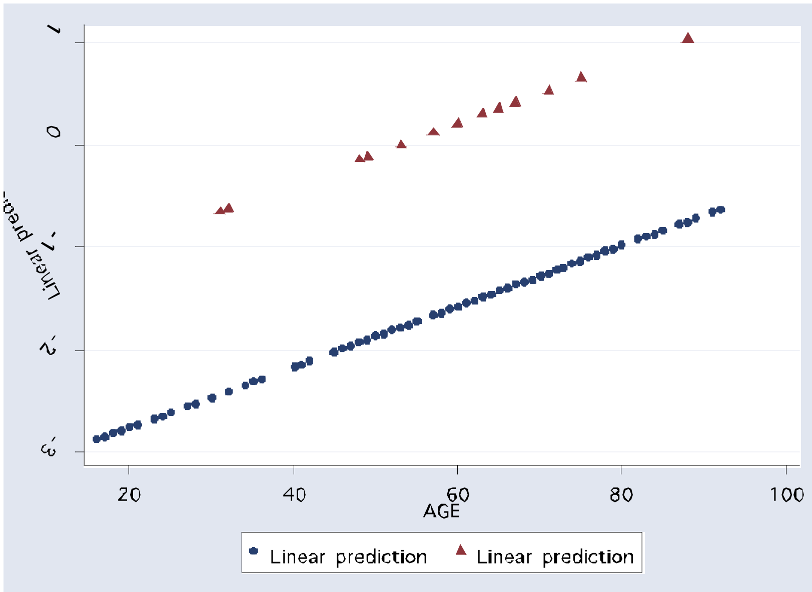
There is a 5.0% increase in the value for the coefficient for CPR when AGE is adjusted for in the model. An increase of this magnitude indicates that AGE probably does not confound the relationship between CPR and STA.

Addition of the interaction of AGE by CPR

```
. gen cprage=CPR*AGE

. logit STA CPR AGE cprage

Logit Estimates                                         Number of obs =    200
                                                        chi2(3)       =  19.05
                                                        Prob > chi2   = 0.0003
Log Likelihood = -90.554825                             Pseudo R2     = 0.0952

------------------------------------------------------------------------------
      STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      CPR |  -3.722935    4.21462     -0.883   0.377      -11.98344    4.537568
      AGE |   .0247665   .0111663      2.218   0.027       .0028809    .0466521
   cprage |   .0941877   .0708038      1.330   0.183      -.0445852    .2329606
    _cons |  -3.041958   .7356889     -4.135   0.000      -4.483882   -1.600034

. predict xbtest4, xb
```

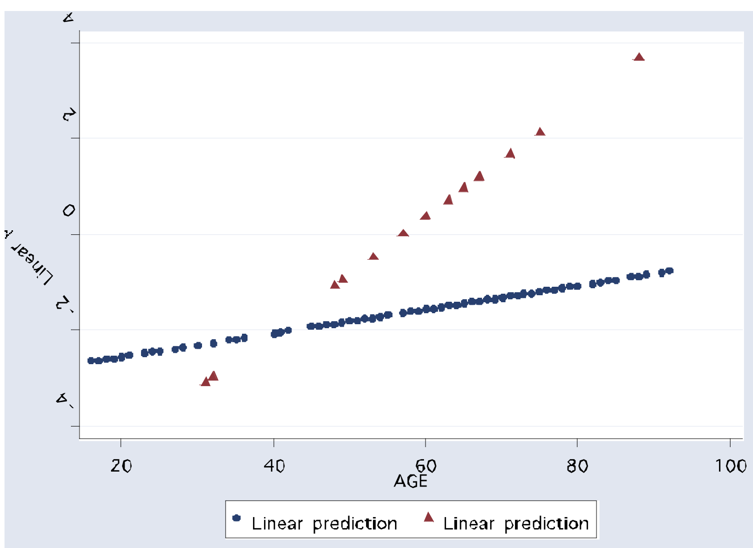
Examine the main effects only and interaction models graphically.

```
. scatter xbtest3 AGE if CPR==0, msymbol(circle) || scatter
xbtest3 AGE if CPR==1, msymbol(triangle)
```



Scatterplot of the main effects model for STA on CPR AGE
logit vs. AGE by CPR
diamond for logit when CRN=0
triangle logit when CRN=1

```
. scatter xbtest4 AGE if CPR==0, msymbol(circle) || scatter
xbtest4 AGE if CPR==1, msymbol(triangle)
```



Scatterplot of the interaction model for STA on CPR AGE CPR*AGE
logit vs. AGE by CPR
diamond for logit when CRN=0
triangle logit when CRN=1

Using the graphical results and any significance tests you feel are needed, select the best model (main effects or interaction) and justify your choice. Estimate the relevant odds ratios.

Model	Constant	CPR	AGE	*CPR*AGE*	log-likelihood	G	p-value
1	-1.540445	1.694596	-	-	-96.11		
2	-3.351956	1.784092	0.0296074	-	-91.98	8.27	0.004
3	-3.041958	-3.722935	0.0247665	0.0941877	-90.55	2.84	0.091

Based on the impression gained from looking at the graph of the logits from the model containing no interaction term, as well as the Wald statistic for the interaction term (cprage) and the results of the likelihood ratio test, it does not appear justified to include the interaction term in the model. There is no effect modification by AGE.

The relevant odds ratio (crude) can be obtained by exponentiating the coefficient for CPR from the original model

$$\widehat{OR} = \exp(1.694596) = 5.44$$

The 95% CI for this odds ratio can be obtained by exponentiating the endpoints of the 95% confidence interval for the coefficient for CPR from the model that does not include AGE:

$$\exp\left(0.5411793\right) \le OR \le \exp\left(2.848012\right)$$

$$1.72 \le OR \le 17.25$$

(f) Consider an analysis for confounding and interaction for the model with STA as the outcome, CAN as the risk factor, and TYP as the potential confounding variable. Perform this analysis using logistic regression modeling and Mantel-Haenszel analysis. Compare the results of the two approaches.

CONFOUNDING

Using logistic regression modeling:

```
. logit STA CAN

Logit Estimates                                        Number of obs =     200
                                                       chi2(1)       =    0.00
                                                       Prob > chi2   =  1.0000
Log Likelihood = -100.08048                            Pseudo R2     =  0.0000

------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     CAN |   1.16e-15   .5892557      0.000   1.000       -1.15492     1.15492
   _cons |  -1.386294    .186339     -7.440   0.000      -1.751512   -1.021077
------------------------------------------------------------------------------

. logit STA CAN TYP

Logit Estimates                                        Number of obs =     200
                                                       chi2(2)       =   18.14
                                                       Prob > chi2   =  0.0001
Log Likelihood = -91.011956                            Pseudo R2     =  0.0906
```

```
------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     CAN |   1.364004   .7817449      1.745   0.081       -.168188    2.896196
     TYP |   2.709722   .8598235      3.151   0.002       1.024499    4.394946
   _cons |  -3.820209   .8563559     -4.461   0.000      -5.498636   -2.141782
------------------------------------------------------------------------------
. logistic STA CAN TYP

Logit Estimates                                        Number of obs =    200
                                                       chi2(2)       =  18.14
                                                       Prob > chi2   = 0.0001
Log Likelihood = -91.011956                            Pseudo R2     = 0.0906

------------------------------------------------------------------------------
     STA | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     CAN |   3.911825   3.058049      1.745   0.081       .8451949    18.10514
     TYP |    15.0251   12.91894      3.151   0.002         2.7857    81.04022
------------------------------------------------------------------------------
```

Using M-H analysis:

```
. by TYP, sort: tabulate STA CAN

-> TYP=        0
           | CAN
       STA |         0          1 |     Total
-----------+----------------------+----------
         0 |        37         14 |        51
         1 |         1          1 |         2
-----------+----------------------+----------
     Total |        38         15 |        53


-> TYP=        1
           | CAN
       STA |         0          1 |     Total
-----------+----------------------+----------
         0 |       107          2 |       109
         1 |        35          3 |        38
-----------+----------------------+----------
     Total |       142          5 |       147
```

$$\widehat{\mathrm{OR}}_{MH} = \frac{\sum_{i=1}^{2} \frac{a_i d_i}{N_i}}{\sum_{i=1}^{2} \frac{b_i c_i}{N_i}}$$

Evaluating the expression to obtain the M-H estimate of the odds ratio:

i	a_i	b_i	c_i	d_i	N_i	$a_i d_i / N_i$	$b_i c_i / N_i$
1	1	1	14	37	53	0.698	0.264
2	3	35	2	107	147	2.184	0.476
Total						2.882	0.740

$$\widehat{\mathrm{OR}}_{MH} = \frac{2.882}{0.740} = 3.895$$

The results of these two approaches are quite similar. After controlling for TYP, the odds of dying prior to discharge from the ICU are nearly four times greater in patients with cancer than in patients without cancer.

INTERACTION

Using logistic regression modeling:

```
. gen cantyp=CAN*TYP

. logit STA CAN TYP cantyp

Logit Estimates                                       Number of obs =    200
                                                      chi2(3)       =  18.24
                                                      Prob > chi2   = 0.0004
Log Likelihood = -90.960895                           Pseudo R2     = 0.0911

------------------------------------------------------------------------------
     STA |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     CAN |   .9718606   1.448604      0.671   0.502      -1.867351    3.811072
     TYP |   2.493437    1.03196      2.416   0.016       .4708326    4.516042
  cantyp |   .5510853   1.723283      0.320   0.749      -2.826487    3.928657
   _cons |  -3.610918   1.013422     -3.563   0.000      -5.597189   -1.624647
------------------------------------------------------------------------------
```

The Wald statistic for the interaction term CANTYP indicates that there is no effect modification of the association between CAN and STA by the variable TYP.

Using M-H analysis to test for heterogeneity across strata:

$$\chi_H^2 = \sum_{i=1}^{2}\left\{w_i\left[\ln\left(\widehat{OR}_i\right)-\ln\left(\widehat{OR}_L\right)\right]^2\right\} \quad \text{where} \quad \widehat{\mathrm{OR}}_L = \exp\left[\frac{\sum w_i \ln\left(\widehat{\mathrm{OR}}_i\right)}{\sum w_i}\right]$$

	TYP=0	TYP=1
$\widehat{\mathrm{OR}}$	2.643	4.586
$[\widehat{\ln(\mathrm{OR})}]$	0.972	1.523
$\widehat{Var}[\widehat{\ln(\mathrm{OR})}]$	2.098	0.871
w	0.477	1.148

$$\widehat{\mathrm{OR}}_L = \exp\left[\frac{0.477(0.972)+1.148(1.523)}{0.477+1.148}\right] = \exp(1.361) = 3.901$$

therefore,

$$\chi_H^2 = \left[(0.477)(0.972-1.361)^2+(1.148)(1.523-1.361)^2\right] = 0.102$$

$$\chi_H^2 \sim \chi^2(1)$$

$$p = 0.749$$

The results of these two approaches are quite similar. There is no indication that TYP is an effect modifier.