Solution to HW9

Problem 6.16

(a) Denote $Y = 1, 2, 3, 4$ for 4 categories of the cholesterol level, $x = 1/0$ for treatment and control, mid-point $y_0 = 1.7, 3.75, 4.5, 5.3$ for baseline cholesterol level. We consider the following cumulative logit model for $\tau_j(x, y_0) = P(Y \le j)$:

$$\text{logit}(\tau_j(x, y_0) = \alpha_j + \beta_1 x + \beta_2 y_0, \quad j = 1, 2, 3.$$

The SAS program and part of the output:

```
data prob6_16;
  input ldl0 c1-c4 t1-t4;
  datalines;
   1.7  18  8  0  0 21  4  2  0
   3.75 16 30 13  2 17 25  6  0
   4.5   0 14 28  7 11 35 36  6
   5.3   0  2 15 22  1  5 14 12
;
data prob6_16; set prob6_16;
    array temp1 {4} c1-c4;
    array temp2 {4} t1-t4;

    do trt=0 to 1;
      do y=1 to 4;
        if trt=0 then
          count=temp1(y);
        else
          count=temp2(y);
        output;
      end;
    end;
    drop c1-c4 t1-t4;
run;

title "Problem 6.16(a)";
proc logistic;
  freq count;
  model y = trt ldl0 / aggregate scale=none;
run;
*********************************************************************
            Score Test for the Proportional Odds Assumption

                  Chi-Square       DF     Pr > ChiSq

                     10.3917        4         0.0343


           Deviance and Pearson Goodness-of-Fit Statistics

    Criterion           Value       DF     Value/DF     Pr > ChiSq

    Deviance          37.0189       19       1.9484         0.0079
    Pearson           41.4759       19       2.1829         0.0021

               Analysis of Maximum Likelihood Estimates

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

Intercept 1     1       4.3071       0.5389       63.8825        <.0001
Intercept 2     1       6.5405       0.5927      121.7646        <.0001
Intercept 3     1       8.7275       0.6493      180.6736        <.0001
trt             1       0.7767       0.2086       13.8669        0.0002
ldl0            1      -1.6029       0.1368      137.3653        <.0001
```

```
                    Odds Ratio Estimates

                  Point          95% Wald
Effect          Estimate      Confidence Limits

trt                2.174       1.445       3.272
ldl0               0.201       0.154       0.263
```

Based on the output, we know that the treatment has a significant effect in reducing cholesterol level: the odds that patients receiving the treatment have better (lower) cholesterol level is 2.17 times the odds that patients receiving the control have better cholesterol level given that they had the same baseline cholesterol level (P-value=0.0002).

(b) Denote D_1, D_2, D_2 three dummy variables for the first three categories for the baseline cholesterol level. Consider the following cumulative logit model for $\tau_j(x, D's) = P(Y \leq j)$:

$$\text{logit}(\tau_j(x, D's) = \alpha_j + \beta_1 x + \beta_2 D_1 + \beta_3 + D_2 \beta_4 D_3, \quad j = 1, 2, 3.$$

The SAS program and part of the output:

```
title "Problem 6.16(b)";
proc logistic;
  class ldl0 / param=ref;
  freq count;
  model y = trt ldl0 / aggregate scale=none;
run;
******************************************************************
            Score Test for the Proportional Odds Assumption

                  Chi-Square       DF     Pr > ChiSq

                      0.8827        8         0.9989


           Deviance and Pearson Goodness-of-Fit Statistics

    Criterion          Value       DF     Value/DF     Pr > ChiSq

    Deviance         14.4679       17       0.8511         0.6338
    Pearson          11.6904       17       0.6877         0.8185

              Analysis of Maximum Likelihood Estimates

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

 Intercept 1        1     -4.9700      0.3721      178.4240       <.0001
 Intercept 2        1     -2.6791      0.3158       71.9755       <.0001
 Intercept 3        1     -0.2062      0.2516        0.6716       0.4125
 trt                1      0.7924      0.2097       14.2765       0.0002
 ldl0      1.7      1      5.6437      0.4661      146.6434       <.0001
 ldl0      3.75     1      3.7689      0.3603      109.4417       <.0001
 ldl0      4.5      1      1.9467      0.3150       38.2025       <.0001

                      Odds Ratio Estimates

                             Point          95% Wald
        Effect             Estimate      Confidence Limits

        trt                   2.209       1.464        3.332
```

Based on the output, we know that the treatment has a significant effect in reducing cholesterol level: the odds that patients receiving the treatment have better cholesterol level is 2.21

times the odds that patients receiving the control have better cholesterol level given that they had the same baseline cholesterol level (P-value=0.0002).

(c) SAS program and output for the CMH test:

```
title "Problem 6.16(c)";
proc freq;
  weight count;
  tables ldl0*trt*y / cmh;
run;
*********************************************************************
                          The FREQ Procedure

                   Summary Statistics for trt by y
                         Controlling for ldl0

     Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

   Statistic    Alternative Hypothesis    DF       Value      Prob
   ---------------------------------------------------------------
       1        Nonzero Correlation        1     15.4001    <.0001
       2        Row Mean Scores Differ     1     15.4001    <.0001
       3        General Association        3     15.4562    0.0015
```

Using score (1,2,3,4) for ending cholesterol level, the CMH test for conditional independence between the treatment and ending cholesterol level given the baseline cholesterol level produces $\chi^2 = 15.4001$, $df = 1$ and $P - value < 0.0001$. Treating the ending cholesterol level as nominal categorical variable, the CMH test for conditional independence between the treatment and ending cholesterol level given the baseline cholesterol level produces $\chi^2 = 15.4562$, $df = 3$ and $P - value < 0.0015$. Both tests indicate that we should reject the conditional independence between the treatment and ending cholesterol level giving baseline cholesterol level.

Problem 8.1

The McNemar test statistic is

$$\chi^2 = \frac{(16-37)^2}{16+37} = 8.32, \quad \mathrm{P-value} = P(\chi_1^2 \geq 8.32) = 0.004.$$

Therefore, we reject the null hypothesis that the diabetes probabilities between MI cases and MI controls are the same. From the table, we know that MI cases have higher diabetes probability than MI controls.

Problem 8.2

(a) Here we can use the McNemar test to test this null hypothesis since we have matched data.

The McNemar test statistic is

$$\chi^2 = \frac{(125-2)^2}{125+2} = 119, \quad \mathrm{P-value} = P(\chi^2_1 \geq 119) = 0.$$

Therefore, we reject the null hypothesis that the population proportions answering "yes" were the same for heaven and hell (almost at any level).

(b) The difference of sample proportions answering "yes" for heaven and hell is:

$$p_1 - p_2 = (833+125)/1120 - (833+2)/1120 = 0.855 - 0.746 = 0.11.$$

The estimated variance (and SE) of the above difference:

$$\widehat{\mathrm{var}}(p_1 - p_2) = \frac{(125+2) - (125-2)^2/1120)}{1120^2} = 9 \times 10^{-5}, \quad \widehat{\mathrm{SE}}(p_1 - p_2) = 0.009.$$

So a 90% CI for the proportion difference is

$$0.11 \pm 1.645 \times 0.009 = [0.095, 0.125].$$

Problem 8.3

(a) Denote Y_{ij} the indicator variable for "yes" for either heaven or hell, x the indicator for heaven (1 for heaven, 0 for hell). The marginal model that will produce marginal odds-ratio is:

$$\mathrm{logit}\{P(Y_{ij} = 1|x)\} = \alpha + x\beta.$$

The correlation among the repeated observations within the same subject can be taken into account using GEE. The SAS program and part of the output is:

```
data prob8_3;
  input heaven hell count;
  datalines;
    1 1 833
    1 0 125
    0 1   2
    0 0 160
  ;

title "recover individual data";
data newdata; set prob8_3;
  retain id;

  if _n_=1 then id=0;

  do i=1 to count;
    id = id+1;
    do x=0 to 1;
      if x=0 then
        y=hell;
      else
        y=heaven;
      output;
    end;
  end;
```

```
run;

title "Problem 8.3(a)";
proc genmod data=newdata descending;
  class id;
  model y = x / dist=bin link=logit;
  repeated subject=id / type=un;
run;
******************************************************************************
                 Analysis Of GEE Parameter Estimates
                  Empirical Standard Error Estimates

                       Standard   95% Confidence
    Parameter Estimate    Error       Limits            Z Pr > |Z|

    Intercept   1.0749   0.0686   0.9405   1.2094   15.67   <.0001
    x           0.7023   0.0621   0.5806   0.8240   11.31   <.0001
```

The odds-ratio estimate is $e^{0.7023} = 2.02$, indicating that the population odds of believing heaven is 2.02 times the population odds of believing hell. The P-value also indicates that these two population odds are not the same.

(b) The conditional model that will produce subject-specific odds-ratio is

$$\text{logit}\{P(Y_{ij} = 1|\alpha_i, x)\} = \alpha_i + x\beta.$$

We use conditional logistic regression to fit the above model. The SAS program and part of the output is:

```
title "Problem 8.3(b)";
proc logistic data=newdata descending;
  model y = x;
  strata id;
run;
********************************************************************
        Analysis of Conditional Maximum Likelihood Estimates

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

 x              1      4.1352       0.7127       33.6606         <.0001

                          Odds Ratio Estimates

                          Point          95% Wald
              Effect    Estimate      Confidence Limits

              x           62.500      15.459     252.677
```

The subject-specific odds-ratio estimate is $e^{4.1352} = 62.5$, indicating for each subject, the odds of of believing heaven is 62.5 times the odds of believing hell. The P-value also indicates that these two subject-specific odds are not the same.