Solution to HW7

Problem 5.3

(a) The LRT comparing model and model 2 is $G^2 = 173.68 - 170.44 = 3.24$ with $df = 155 - 152 = 3$. The P-value is $P(\chi^2_3 \geq 3.24) = 0.356$, which suggests that we can remove the three-factor term from model 1.

(b) The LRT comparing model 3a to model 2 is $G^2_{3a} = 177.34 - 173.68 = 3.66$ with $df = 158 - 155 = 3$ and P-value $= P(\chi^2_3 \geq 3.66) = 0.3$. The LRT comparing model 3b to model 2 is $G^2_{3b} = 181.56.34 - 173.68 = 7.88$ with $df = 161 - 155 = 6$ and P-value $= P(\chi^2_6 \geq 7.88) = 0.247$. The LRT comparing model 3c to model 2 is $G^2_{3c} = 173.69 - 173.68 = 0.01$ with $df = 157 - 155 = 2$ and P-value $= = P(\chi^2_2 \geq 0.01) = 0.995$. Obviously, model 3c fits the data equally well as model 2. Therefore, we should select model 3c.

(c) The LRT comparing model 4a to model 3c is $G^2 = 181.64 - 173.69 \approx 8$ with $df = 163 - 157 = 6$ and P-value $= P(\chi^2_6 \geq 8) = 0.24$. LRT comparing model 4b to model 3c is $G^2 = 177.61 - 173.69 = 3.92$ with $df = 160 - 157 = 3$ and P-value $= P(\chi^2_3 \geq 3.92) = 0.27$. Therefore, model 4b is slightly preferred over model 4a.

(d) If we pick model 4b at (c), then the LRT comparing model 5 to model 4b is is $G^2 = 186.61 - 177.61 = 9$ with $df = 166 - 160 = 6$ and P-value $= P(\chi^2_6 \geq 9) = 0.174$. The P-value indicates that we can use this simplified main-effects only model.

(e) If we are using AIC as the model selection criterion, model 5 is preferred.

Problem 5.7

(a) Denote by T, J, E, S dummy variables for *thinking, judging, extroversion* and *sensing*, and π the smoking probability. Then the four models are:

(1) Intercept only model:

$$\text{logit}(\pi) = \alpha.$$

The number of model parameters is 1. $AIC = 1130 + 2 \times 1 = 1132$.

(2) Main effects only model:

$$\text{logit}(\pi) = \alpha + \beta_1 T + \beta_2 J + \beta_3 E + \beta_4 S.$$

The number of model parameters is 5. $AIC = 1124.86 + 2 \times 5 = 1134.86$.

(3) Model with all two-way interactions:

$$\begin{aligned}\text{logit}(\pi) &= \alpha + \beta_1 T + \beta_2 J + \beta_3 E + \beta_4 S \\ &\quad + \beta_{12} T \times J + \beta_{13} T \times E + \beta_{14} T \times S + \beta_{23} J \times E + \beta_{24} J \times S + \beta_{34} E \times S.\end{aligned}$$

The number of model parameters is 11. $AIC = 1119.87 + 2 \times 11 = 1141.87$. (4) Model with all three-way interactions:

$$\begin{aligned}\text{logit}(\pi) &= \alpha + \beta_1 T + \beta_2 J + \beta_3 E + \beta_4 S \\ &\quad + \beta_{12} T \times J + \beta_{13} T \times E + \beta_{14} T \times S + \beta_{23} J \times E + \beta_{24} J \times S + \beta_{34} E \times S \\ &\quad + \beta_{123} T \times J \times E + \beta_{124} T \times J \times S + \beta_{134} T \times E \times S + \beta_{234} J \times E \times S\end{aligned}$$

The number of model parameters is 15. $AIC = 1116.47 + 2 \times 15 = 1146.47$.

(b) The model with the smallest AIC among previous 4 models is the model with intercept only. Therefore, the intercept only model is preferable according to AIC, which indicates that frequent smoking is independent of those personality types.

(c) For a good diagnostic test, the difference between sensitivity and one minus specificity should be high. However, for the classification in this problem, the difference is 0.48 - (1-0.55) = 0.03, a value very close to 0. This means that the classification is not much better than a random guess using probability 0.5. The area under the ROC curve $c = 0.55$ is also not much greater than 0.5, the area under the ROC curve from the intercept only model (probability of being frequent smoker is independent of personality). Therefore, the knowledge of personality does not predict well whether or not an individual is a frequent smoker.

Problem 5.10

(a) The SAS program and relevant output are:

```
data crab;
input color spine width satell weight;
   weight=weight/1000; color=color-1;
   y=(satell>0);
datalines;
3  3  28.3  8  3050
4  3  22.5  0  1550
2  1  26.0  9  2300
4  3  24.8  0  2100
.
.
.
;
```

```
title "Logit model for the prob of having satellites with predictor wt";
proc logistic data=crab descending;
  model y=weight / lackfit outroc=roc;
  output out=out predicted=pihat lower=lower upper=upper / alpha=0.05;
run;

data out; set out;
  if pihat>0.642 then yhat=1;
  else yhat=0;
run;

proc freq;
  tables y*yhat / nocol nopercent;
run;

options ls=70 ps=25;

Title "ROC Curve";
proc plot data=roc;
  plot _sensit_*_1mspec_;
run;
```

```
*********************************************************************
                         Model Fit Statistics

                                                 Intercept
                               Intercept               and
                Criterion           Only        Covariates

                AIC              227.759           199.737
                SC               230.912           206.044
                -2 Log L         225.759           195.737

              Analysis of Maximum Likelihood Estimates

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

 Intercept     1     -3.6947       0.8802       17.6196        <.0001
 weight        1      1.8151       0.3767       23.2183        <.0001

    Association of Predicted Probabilities and Observed Responses

          Percent Concordant     72.7    Somers' D    0.476
          Percent Discordant     25.1    Gamma        0.487
          Percent Tied            2.2    Tau-a        0.220
          Pairs                  6882    c            0.738

              Hosmer and Lemeshow Goodness-of-Fit Test

                  Chi-Square       DF     Pr > ChiSq

                     12.6818        8         0.1233

                         Table of y by yhat

                 y         yhat

                 Frequency|
                 Row Pct  |       0|       1|  Total
                 ---------+--------+--------+
                        0 |     45 |     17 |     62
                          |  72.58 |  27.42 |
                 ---------+--------+--------+
                        1 |     43 |     68 |    111
                          |  38.74 |  61.26 |
                 ---------+--------+--------+
```

So the sensitivity of this classification is 61.26% and the specificity is 72.58%. The classification is reasonably good since sensitivity - (1-specificity) = 0.6126 - (1 - 0.7258) = 0.34.

(b) The ROC curve looks like:

```
                              ROC Curve                              3

  Plot of _SENSIT_*_1MSPEC_.  Legend: A = 1 obs, B = 2 obs, etc.

           |
  0.9730 +                                         A    AAA ABA AA
S          |                                   C    A
e          |                               C
n          |                         AA  B  A
s 0.6486 +                     AA  B
i          |                 BA
t          |           A A A A
i          |          AA
v 0.3243 +           CC
i          |         B
t          |        C
y          |   F
  0.0000 +   D
           |
           ---+----------+----------+----------+----------+----------+--
             0.0        0.2        0.4        0.6        0.8        1.0

                               1 - Specificity
```

The area under the ROC is 0.738, which also measures Y_i and $\widehat{Y_i}$ being concordant. This value is somewhat large, indicating reasonable fit of the model to the data.

(c) The Hosmer and Lemeshow goodness-of-fit test produces $\chi^2 = 12.6818$ with $df = 8$, reasonable fit (P-value=0.1233).

(d) Logit model with x and x^2:

```
title "Logit model for the prob of having satellites with wt and wt^2";
proc logistic data=crab descending;
  model y=weight weight*weight;
run;
```

```
***********************************************************************
                         Model Fit Statistics

                                             Intercept
                              Intercept            and
                Criterion          Only     Covariates

                AIC             227.759        201.460
                SC              230.912        210.920
                -2 Log L        225.759        195.460


               Testing Global Null Hypothesis: BETA=0

       Test                 Chi-Square       DF     Pr > ChiSq

       Likelihood Ratio        30.2981        2         <.0001
       Score                   27.4121        2         <.0001
       Wald                    22.4531        2         <.0001


               Analysis of Maximum Likelihood Estimates

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

  Intercept        1     -1.8877        3.5494        0.2829        0.5948
  weight           1      0.2182        3.0818        0.0050        0.9436
  weight*weight    1      0.3393        0.6543        0.2689        0.6041

    Association of Predicted Probabilities and Observed Responses

          Percent Concordant     72.7    Somers' D    0.476
          Percent Discordant     25.1    Gamma        0.487
          Percent Tied            2.2    Tau-a        0.220
          Pairs                  6882    c            0.738
```

In the model with x only, this predictor is significant (P-value < 0.0001). However, in the model with x and x^2, each term is not significant, even though both terms together are significant (P-values of LRT, score and Wald are all less than 0.0001). This indicates that x and x^2 in the data range are highly correlated (Pearson correlation coefficient $= 0.98$). Also the model with x and x^2 basically has the same fit as the model with x only since they have the same c-index (0.738).

(e) Adding x^2 to the model with x only does not improve the fit (The -2L became 195.460 from 195.737). The AIC for the model with x only is 199.737, and the AIC for the model with x and x^2 is 201.460. Therefore, the model with x only is preferred.

Problem 5.18

We use the following SAS program to answer the questions in this problem:

```
data prob5_18;
  input city $ smoke $ y y0 @@;
  n = y+y0; dsmoke = (smoke="Yes");
  datalines;
Beijing  Yes 126 100 Harbin    Yes 402 308
Beijing  No  35  61  Harbin    No  121 215
Shanghai Yes 908 688 Zhengzhou Yes 182 156
Shanghai No  497 807 Zhengzhou No  72  98
Shenyang Yes 913 747 Taiyuan   Yes 60  99
Shenyang No  336 598 Taiyuan   No  11  43
Nanjing  Yes 235 172 Nanchang  Yes 104 89
Nanjing  No  58  121 Nanchang  No  21  36
;

proc genmod;
  class city;
  model y/n=city dsmoke / noint dist=bin link=logit residuals;
run;
```

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

                Criteria For Assessing Goodness Of Fit

   Criterion                     DF           Value        Value/DF

   Deviance                       7          5.1958          0.7423
   Scaled Deviance                7          5.1958          0.7423
   Pearson Chi-Square             7          5.1999          0.7428
   Scaled Pearson X2              7          5.1999          0.7428

         Analysis Of Maximum Likelihood Parameter Estimates

                                          Standard   Wald 95% Confidence
Parameter              DF   Estimate       Error          Limits

Intercept               0     0.0000      0.0000      0.0000      0.0000
city        Beijing     1    -0.5487      0.1180     -0.7800     -0.3174
city        Harbin      1    -0.5305      0.0707     -0.6690     -0.3920
city        Nanchang    1    -0.6036      0.1333     -0.8648     -0.3423
city        Nanjing     1    -0.5429      0.0902     -0.7197     -0.3661
city        Shanghai    1    -0.4931      0.0460     -0.5833     -0.4028
city        Shenyang    1    -0.5764      0.0504     -0.6752     -0.4776
city        Taiyuan     1    -1.2944      0.1517     -1.5916     -0.9971
city        Zhengzho    1    -0.5199      0.0956     -0.7073     -0.3325
dsmoke                  1     0.7771      0.0468      0.6854      0.8687
Scale                   0     1.0000      0.0000      1.0000      1.0000
```

Observation Statistics

Observation	Raw Residual / Std Deviance Residual	Pearson Residual / Std Pearson Residual	Deviance Residual / Likelihood Residual
1	0.1523411	0.0203994	0.0204005
	0.0388654	0.0388633	0.0388639
2	3.4548523	0.2612897	0.2613975
	0.5004287	0.5002224	0.5002787
3	-0.152342	-0.032274	-0.032284
	-0.038875	-0.038864	-0.038872
4	-3.454855	-0.390294	-0.391056
	-0.501199	-0.500223	-0.500818
5	-2.559635	-0.129436	-0.129416
	-0.247103	-0.247141	-0.24713
6	-8.611169	-0.944526	-0.942657
	-1.708292	-1.711679	-1.710649
7	2.5596317	0.1460948	0.146046
	0.2470577	0.2471403	0.2471115
8	8.6111605	1.3657726	1.3547009
	1.6978015	1.7116774	1.7028562
9	0.0124937	0.0006164	0.0006164
	0.0012648	0.0012648	0.0012648
10	0.6162013	0.1010234	0.1009535
	0.2293976	0.2295564	0.2295257
11	-0.0125	-0.000852	-0.000852
	-0.001265	-0.001265	-0.001265
12	-0.616204	-0.204075	-0.20542
	-0.23107	-0.229557	-0.230754
13	7.7841102	0.7769824	0.7782543
	1.4862281	1.4837992	1.4844656
14	-0.849177	-0.12271	-0.12268
	-0.26831	-0.268376	-0.268363
15	-7.784123	-1.206763	-1.21781
	-1.497385	-1.483802	-1.4928
16	0.8491764	0.2352746	0.2345504
	0.26755	0.2683762	0.2677415

(a) The smoking effect: $\widehat{\beta} = 0.7771$, $e^{0.7771} = 2.18$, indicating that at any one of those Chinese cities in the study, the odds of developing a lung caner among smokers is 2.18 times the odds of developing a lung caner among non-smokers. Since the lung cancer is a rare disease, we can say that at any one of those Chinese cities in the study, smokers are 118% more likely to develop a lung cancer than non-smokers.

(b) Here we can use the Pearson χ^2 or deviance for goodness-of-fit of the model since we have true binomial data with very large binomial sample sizes. The Pearson χ^2 goodness-of-fit statistic for the model in (a) is $\chi^2 = 5.2$ with $df = 7$. The P-value of the test is $P(\chi^2_7 > 5.2) = 0.64$, indicating a very good fit.

(c) All standardized Pearson residuals are within (-2, 2), indicating no outliers.

Problem 5.23

(a) Denote by d_k the dummy variable for level $k (k = 1, 2, ..., 5)$. The logit model for the cured probability becomes:

$$\text{logit}(\pi) = \beta x + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 d_4 + \beta_5 d_5.$$

The cured probability for level one is:

$$\pi_1(x) = \frac{e^{\beta x+\beta_1}}{1+e^{\beta x+\beta_1}}, x = 0, 1$$

and the likelihood contributed from data in level 1 is:

$$\left(\frac{1}{1+e^{\beta_1}}\right)^6 \left(\frac{1}{1+e^{\beta+\beta_1}}\right)^5$$

For a given β, $\beta_1 \to -\infty$ will maximize the above likelihood.

The cured probability for level five is:

$$\pi_5(x) = \frac{e^{\beta x+\beta_5}}{1+e^{\beta x+\beta_5}}, x = 0, 1$$

and the likelihood contributed from data in level 5 is:

$$\left(\frac{e^{\beta_5}}{1+e^{\beta_5}}\right)^2 \left(\frac{e^{\beta+\beta_5}}{1+e^{\beta+\beta_5}}\right)^5.$$

For a given β, $\beta_5 \to \infty$ will maximize the above likelihood.

The SAS program for the above model and part of the output are:

```
data prob5_23;
  input level $ delay y y0;
  d1 = (level="1/8");
  d2 = (level="1/4");
  d3 = (level="1/2");
  d4 = (level="1");
  d5 = (level="4");
  n = y+y0;
  datalines;
    1/8 0 0 6
    1/8 1 0 5
    1/4 0 3 3
    1/4 1 0 6
    1/2 0 6 0
    1/2 1 2 4
    1   0 5 1
    1   1 6 0
    4   0 2 0
    4   1 5 0
    ;

proc genmod ;
  model y/n = delay d1 d2 d3 d4 d5 / noint dist=bin link=logit type3;
run;

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

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept	0	0.0000	0.0000	0.0000	0.0000	.
delay	1	-2.5496	1.1752	-4.8530	-0.2463	4.71
d1	1	-27.1660	313505.2	-614486	614431.8	0.00
d2	1	-0.2339	0.7737	-1.7503	1.2825	0.09
d3	1	2.2592	1.1236	0.0569	4.4614	4.04
d4	1	4.2626	1.5146	1.2942	7.2311	7.92
d5	1	29.2763	280184.5	-549122	549180.9	0.00
Scale	0	1.0000	0.0000	1.0000	1.0000	

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
delay	1	6.80	0.0091
d1	1	8.94	0.0028
d2	1	0.09	0.7615
d3	1	6.36	0.0116
d4	1	16.46	<.0001
d5	1	16.41	<.0001

From the output, we see that $\widehat{\beta}_1 = -27.166$ and $\widehat{\beta}_5 = 29.28$, consistent to the above theoretical result.

(b) Conditional independence of X and Y given levels is equivalent to $H_0 : \beta = 0$ (β is the coefficient of x in the logit model). The LRT stat is $G^2 = 6.8$ with $df = 1$, so the P-value = 0.0091.

(c) The XY conditional odds-ratio is estimated as $e^{-2.5496} = 0.078$. Interpretation: At any penicillin level, the odds of rabbits being cured with immediate injection is 0.078 times the cure odds with $1\frac{1}{2}$ hour delay.

(d) The SAS program and output for conditional logistic regression is

```
title "Conditional logist treating level as nuisance";
proc logistic;
  class level;
  model y/n = delay;
  strata level;
run;
```

```
*****************************************************************************
        Analysis of Conditional Maximum Likelihood Estimates

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

 delay         1     -2.3381       1.1293        4.2862        0.0384
```

The XY conditional odds-ratio is estimated as $e^{-2.3381} = 0.0097$, very similar to the one obtained in (c).

(e) The large-sample CMH test for XY conditional independence given level:

$$\chi^2 = \frac{\{(3 - 3 \times 6/12) + (6 - 6 \times 8/12) + (5 - 11 \times 6/12)\}^2}{3 \times 9 \times 6 \times 6/(11 \times 12^2) + 8 \times 4 \times 6 \times 6/(11 \times 12^2) + 11 \times 1 \times 6 \times 6/(11 \times 12^2)} = 5.6571,$$

with $df = 1$. So the P-value is $P(\chi_1^2 \geq 5.6571) = 0.0174$. Therefore, we reject the conditional independence of X and Y given penicillin levels using the large-sample CMH test.

The SAS program for conducting exact CMH test and part of the output are

```
title "Exact CMH test";
proc logistic;
  class level / param=ref;
  model y/n = delay level;
  exact delay;
run;
```

```
****************************************************************************
                        The LOGISTIC Procedure
                      Exact Conditional Analysis
                       Exact Conditional Tests

                                              --- p-Value ---
          Effect   Test          Statistic    Exact       Mid

          delay    Score            5.6571   0.0399    0.0306
                   Probability      0.0186   0.0399    0.0306
```

The exact P-value = 0.0399 and the exact mid P-value = 0.0306. Therefore there is a strong evidence to reject the conditional independence of X and Y given penicillin levels.