STAT 659 Spring 2016
Homework 8 Solution

5.13

The Hosmer-Lemeshow test statistic is 6.7503 which has a chi-square distribution with degree of freedom 8. The P-value is $0.5638 > 0.05$ which implies the model is adequate.

5.9

(a) No, the deviance is only suitable for the categorical predictors. This data has small number of observations for unique profile, the deviance do not have approximate chi-squared distribution.
(d) Since the p-value is $0.3594 > 0.05$, the model seems adequate.

5.10

(c)The Hosmer-Lemeshow test statistic is 12.6818 which has a chi-square distribution with degree of freedom 8. The P-value is $0.1233 > 0.05$ which implies the model is adequate.

5.14

(a) For both grouped and ungrouped data, $-2L_0 = 16.301$, $-2L_1 = 11.028$ which are not dependent on the form of data entry.

(b) For ungrouped data, $G^2(M_0) = 16.301, G^2(M_1) = 11.028$ while for grouped data, $G^2(M_0) = 6.2568, G^2(M_1) = 0.9844$. So we can see the deviance are dependent on the form of data entry.

(c) $G^2(M_0|M_1) = -2(L_0 - L_s) + 2(L_1 - L_s) = -2(L_0 - L_1)$. Since L_0, L_1 do not dependent on the data form, then $G^2(M_0|M_1)$ does not depend on the data form.

5.16

Let gender $= \begin{cases} 1 & \text{females} \\ 0 & \text{males} \end{cases}$ and race $= \begin{cases} 1 & \text{white} \\ 0 & \text{black} \end{cases}$. Then the fitted model is

$$\text{logit}\pi = -0.85405 + 0.35244\text{gender} + 1.01547\text{race}$$

Further test shows that the interaction term is not significant. The goodness of fit test statistics is 0.00012 with P-value 0.9913, so this model is adequate.

5.18

(a) The fitted model is $\text{logit}\pi = -0.548682 + 0.777062\text{smoke} + 0.0556\text{Shanghai} - 0.0277\text{Shenyang} + 0.0058\text{Nanjing} + 0.0182\text{Harbin} + 0.0288\text{Zhangzhou} - 0.7457\text{Taiyuan} - 0.0549\text{Nanchang}$. The odds ratio between smoker and nonsmoker is $\exp(0.777062) = 2.1751$.

(b) The Pearson's Chi-square statistic is $5.1998 \sim \chi^2_7$ with P-value 0.6356. So the model is adequate.

(c) None of the Pearsons residuals or Deviance are greater than 2 or less than -2 and the residual plots have no patterns, which indicate that the model fit is adequate.

5.19

(a) If we let $\text{dept}_\text{i} = \begin{cases} 1 & \text{admitted to school i} \\ 0 & \text{Otherwise} \end{cases}, i = 1, 2, \cdots, 5$, then the model is $\text{logit}\pi = \alpha + \beta_1\text{dept}_1 + \cdots + \beta_5\text{dept}_5$.

(b) The P-value is $0.00137 < 0.05$, so the model fit is not adequate.

(c) The model is lack of fit for the data of department one and the residual is too large.

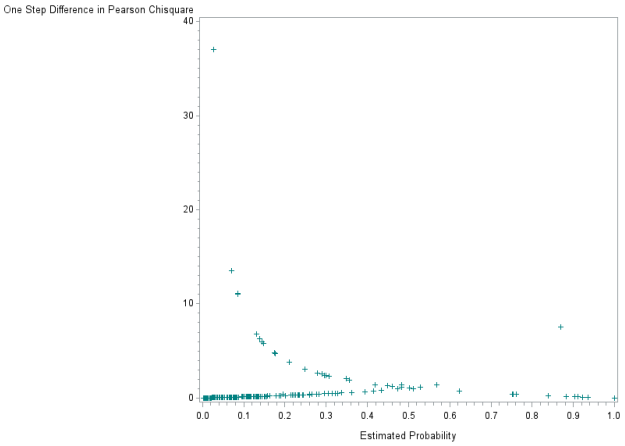
(d) The residual is -4.15, which means there are fewer males admitted in department one than expected.

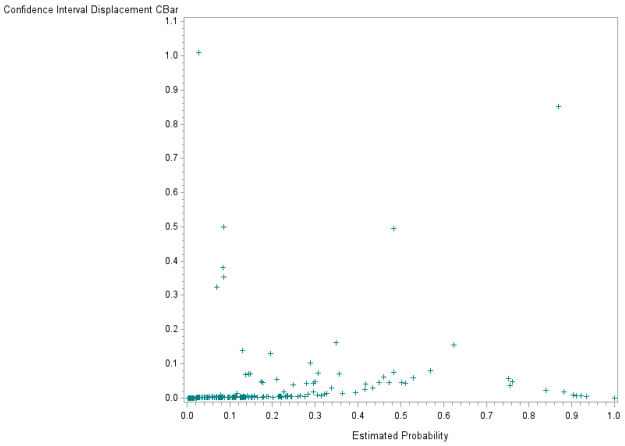
(e) This is Simpson's paradox. This is because although conditional on each department, the admitted rate for women is higher than men, the men apply in larger number for the department which has high acceptance rate.

Variable Selection Problem (continued):

4.Only the variables age, sys, typ, and loc_2 appear to be statistically. Next, let's check for interactions. There are 6 possible two-way interactions among the 4 explanatory variables. We checked the results of adding one interaction at a time to the main effects model. The only interactions that were significant were age*sys. Thus the final model is sta=age sys typ loc age*sys.
5. There is a quasi-complete separation problem, if the above model is fitted. This problem arises due to an empty cell with loc=1 and sta=0. One way to avoid the problem is use the loc value which is set to one if the original value is one or two. The Hosmer-Lemeshow goodness of fit test statistic is 7.2348 with P-value 0.515 which indicates that the fitted model is adequate. Then we produce plots for diagnosis of influence,such as the change of chi-square test statistic for deleting each data and the CI displacement c. From these plots

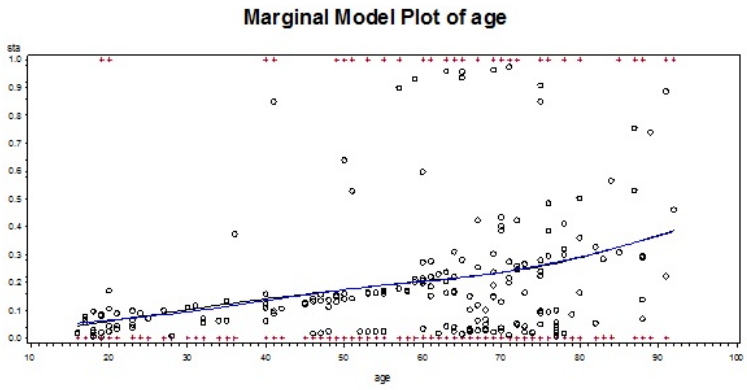
we can find three points having large influence.

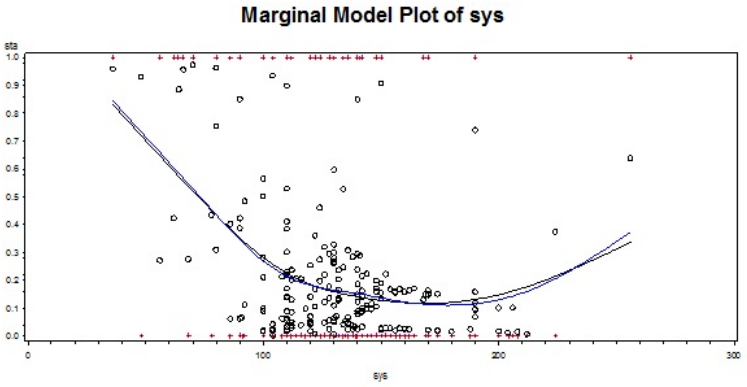




Obs	id	sta	age	sys	typ	loc	pred	hat	deltaci	deltadev	deltachi
1	18	0	59	48	1	1	0.93321	0.07341	1.10686	6.51921	15.0787
2	151	0	89	190	1	1	0.70720	0.27862	0.93284	3.38935	3.3481
3	172	1	70	168	0	1	0.40190	0.27628	0.56811	2.39124	2.0563
4	170	1	75	130	0	0	0.04300	0.02414	0.55055	6.84355	22.8050
5	177	1	40	86	1	0	0.06928	0.03377	0.46962	5.80894	13.9047

6. Since the two fits are similar for both marginal model plots, that the linearity of age and sys is appropriate.





(Only for students having taken STAT 414, 610 or STAT 630) 5.25

We use LI to denote Lymphocytic Infiltration and set it to be one if the level is high and zero otherwise. Then let OP $= \begin{cases} 1 & \text{Osteoblastic Pathology is yes} \\ 0 & \text{Otherwise} \end{cases}$ and sex $= \begin{cases} 1 & \text{if female} \\ 0 & \text{Otherwise} \end{cases}$.

(a) The likelihood ratio test statistic for LI is 10.8071 with P-value 0.0010, so LI effect is significant; the likelihood ratio test statistic for sex is 5.8795 with P-value 0.0315, so sex effect is significant; the likelihood ratio test statistic for OP is 5.5349 with P-value 0.0186, thus the OP effect is also significant.

(b) The main effect model is $\text{logit}\pi = 0.4715 + 12.6349\text{LI} + 1.6362\text{sex} - 1.2204\text{OP}$. The standard deviation for LI is 230.1 which is huge. This is because the responses are all one for high level and the $\text{logit}\pi = +\infty$ for high level. Then the coefficient of LI should be positive infinity to obtain a good fit.

(c) For the conditional exact test for LI, the score test statistic is 4.5416 with P-value $0.0606 > 0.05$. So the LI effect is not significant.

(d) The 95 percent confidence interval for LI effect is $(-0.1615, +\infty)$.