Homework # 4, Stat 526

1. The data reports the number of high school graduates(with Yes or No) by race(White, Black, and Hispanic), gender (male and female), and family structure (intact and nonintact) from a subsample of young adults aged from 25 to 30 from the National Longitudinal Survey of Youth(NSLY). The research wants to know how the probability of "yes" relates the other three variables.

   (a) Fit an additive logistic linear model and test the goodness of the fit, and test the significant of all the interaction effects jointly.

   (b) Test the significance of each of the three variables and write down your final model.

   (c) Based on your final model. Look at the qq-plot of the Pearson residuals. Is there an outlier in this dataset. Compute the estimate of the dispersion parameter and justify whether there is a significant over-dispersion in this data set.

   (d) Compute the 95% confidence interval for the probability for a male Hispanic person in the intact family.

2. The dataset reports the frequency for the job satisfaction and Income. The level of Job Satisfaction has four levels: Very Dissatisfied(1), Little Dissatisfied(2), Moderately Satisfied(3), and Very Satisfied(4); and the Income has four levels: $< 6000(1)$, $6000 - 15000(2)$, $15000 - 25000(3)$, $> 25000(4)$.

   (a) Fit an independent Poisson model based on the frequency as the response and the other two as **factor** predictors. Is the fit good or bad. What can you say about the relationship between the predictors in this part(independent or not).
   Hint: In R, you should use function *factor* in front of the two predictors. Otherwise, R will treat them as continuous variables.

   (b) Define the interaction term as the product of the two variables and treat is as a continuous variables. Fit a Poisson model with the main effect as factor variables and the interaction effect as a continuous variable. Test the signifiance of the interaction effect. What can you say about the relationship between the predictors in this part(independent or not).

   (c) Is there an inconsistence between these two parts.

3. In a success/failure experiment, the probability of success may depend on an independent variable x. Suppose 5 values of x were considered and each of them were tested 30 times. Therefore, we have the following summarized table

x	Success	Failure
1	6	24
2	16	14
3	16	14
4	24	6
5	27	3

(a) Fit a logistic regression model, and compute the deviance and Pearson goodness of fit. Explain.

(b) Consider all the three link function for binomial distribution. Compute the 95% confidence interval for the proportion when $x = 2.5$.

(c) Suppose the original data is given by the index of success of the 150 items (1 is success and 0 is failure). Fit a Bernoulli logistic regression model. Compare the estimate of parameters and the goodness of fit statistics with the logistic regression model in part (a). Explain the reason for the difference.