Second Assignment: Generalized linear models

1. (1.5 points) A marketing firm is investigating the likelihood that a family in a certain geographic region will buy a new car within the next year. A random sample of 33 families from this region was selected. A follow-up interview 12 months later was conducted to record the annual family income (X_1, in thousand dollars), and whether the family purchased a new car ($Y = 1$) or not ($Y = 0$). The output of a statistical model used to analyze the data is given below

```
glm(formula = Y ~ income, family = binomial(link = "logit"), data = insurance)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.98079    0.85720  -2.311   0.0208 *
income       0.04342    0.02011   2.159   0.0308 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 39.305  on 31  degrees of freedom
AIC: 43.305

> summary(fit)$cov.unscaled
            (Intercept)        income
(Intercept)  0.73478346 -0.0153955424
income      -0.01539554  0.0004044087
```

   - Write the model, and interpret the parameters
   - Calculated a 95% confidence interval for the probability that a family with annual income of 60 thousand dollars will purchase a new car next year.
   - The analyst considers grouping the individuals into 6 levels of income, calculating the number of car purchases per income group, and fitting the same logistic regression to the grouped data. What would be the test used to check the goodness of fit of the model and what would be the degrees of freedom of the test statistic?.

2. (2.5 points) Low birth weight, defined as birth weight less than 2500 grams, is an outcome that has been of concern to physicians for years. This is because of the fact that infant mortality rates and birth defect

rates are higher for low birth weight babies. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term, and, consequently, of delivering a baby of normal birth weight.
Data were collected as part of a larger study at Baystate Medical Center in Springfield, Massachusetts. This data set contains information on 189 births to women seen in the obstetrics clinic. Fifty-nine of these births were low birth weight. The variables identified in the code sheet given in table bellow have been shown to be associated with low birth weight in the obstetrical literature. The goal of the current study was to determine whether these variables were risk factors in the clinic population being served by Baystate Medical Center. Actual observed variable values have been modified to protect subject confidentiality. The data are available in the dataset `birthwt` which is part of the package `MASS`

```
-----------------------------------------------------------------
low     Birth weight      (0 = Peso $>$ 2500g,
l = Peso $<$ 2500g)
age     Mother's age (in years)
lwt     Weight of the mother before pregnancy
race    Race (1 = White, 2 = Black, 3 = Other)
smoke   Smoker during pregnancy (1 = Yes, 0 = No)
-----------------------------------------------------------------
```

- Fit the best model using AIC and LRT (check for interactions).
- Use the Hosman-Lemeshow test to check the goodness of fit of the final model (using 10 groups).
- Check the model assumptions using residual plots
- What is the total error rate of the model, is it a good predictive model?
- What are the characteristics of the mothers with higher probability of having babies with low birth weight?, what is the characteristic that has the highest mpact on the predicted probability?

3. (1.5 point) Fit the best logistic regression model to predict the probability self-perceived health using the predictors *sex* and *weight*. Use the LRT to test if the terms in the model are significant.

- Interpret the coefficients in terms of odds ratios

- Plot the predicted probabilities for males and females
- Given that a person has $weight = 75kg$, what is the relative risk and odds ratio of self-perceived good health of a female compared with a male?
- Calculate the estimated expected probability of females of 70kg and 110kg and give a confidence interval for the prediction

4. (1 point) Use all predictors available in the dataset *health* to find the best subset of predictors (and their possible interactions) using LRT, AIC and BIC. Are the chosen models the same?. If the answer is not, which one would you use as your final model?. Check the predictive accuracy of the final model.

5. (1 point) Find the best model for the property crime rates used in chapter 6 and interpret the parameters

6. (2.5 points) The aim of this task is to examine the relationship between the number of physician office visits for a person (ofp) and a set of explanatory variables for individuals on Medicare. Their data are contained in the file dt.csv.

   The explanatory variables are number of hospital stays (*hosp*), number of chronic conditions (*numchron*), gender (*gender*; male = 1, female = 0), number of years of education (*school*), private insurance (*privins*; yes = 1, no = 0), $health_excellent$ and $health_poor$, these two are self-perceived health status indicators that take on a value of yes = 1 or no = 0, and they cannot both be 1 (both equal to 0 indicates ?average? health). Using these data, complete the following:

   - Estimate the Poisson regression model to predict the number of physician office visits and interpret the coefficients
   - Compare the number of zero-visit counts in the data to the number predicted by the model and comment. Can you think of a possible explanation for why there are so many zeroes in the data?
   - Estimate the zero-inflated Poisson regression model to predict the number of physician office visits. Use all of the explanatory variables for the $log(\mu)$ part of the model and no explanatory variables in the ϕ part of the model. Interpret the model fit results
   - Do the previous item again, but now use all of the explanatory variables to estimate ϕ. Interpret the model fit results and compare this model to the previous ZIP model using a LRT
   - Examine how well each model estimates the number of 0 counts