Stat 6021: Homework Set 10

1. For this question, you will use a data set `Boston`, which comes from the `MASS` package in R. This is the same data set that you saw in the live session. The purpose of this question is to classify a town as being a high- or low-crime town based on some predictors. The governor of Massachusetts is most interested in identifying towns that have a high crime rate.

   (a) Before fitting your logistic regression model you will need to create a new variable. The variable *crim* is the per capita crime rate of the town. Create a new variable that classifies *crim* in the following manner. Define a town to have a

   - low crime rate, if its crime rate is less than 1 per capita.
   - high crime rate, otherwise

   Using the `ifelse()` function will be very helpful for this.

   I highly recommend to add this newly created variable to the existing data frame. Also, use the `contrasts()` function to see which class is the reference class for this newly created variable. What is the reference class?

   (b) Randomly split the data into a testing and training set of equal sizes. For consistency of results among all groups, use `set.seed(199)`. Next, using the training set, fit a logistic regression model, with your newly created variable as the binary response variable, and with the following predictors: *indus*, *nox*, *rad*, *tax*, *lstat*, and *medv*. Then validate your model on the test data by creating an ROC curve. What does your ROC curve tell you?

   (c) Find the AUC associated with your ROC curve. What does your AUC tell you?

   (d) Create a confusion matrix using a cutoff of 0.5. What is the false positive rate? What is the false negative rate? Note: Be very careful with the coding associated with your response variable. You may want to use the `levels()` function to check how R was coding your response variable.

   (e) Bearing in mind the governor is most interested in identifying towns with high crime rates, how would you adjust the cutoff value from 0.5? Briefly explain why.

2. (No R required) A study was undertaken to determine the association between several predictors and the duration of pregnancies. The response variable, pregnancy duration, was recorded as a three-class variable: pre-term for pregnancies lasting less than 36 weeks, intermediate term for pregnancies lasting between 36 and 37 weeks, and full term for pregnancies lasting more than 37 weeks. The predictors are:

   - *nutrition*: index of nutritional status, higher scores denote better nutritional status
   - *less20*: 1 =less than 20 years old, 0 otherwise
   - *greater30*: 1 =greater than 30 years old, 0 otherwise
   - *alcohol*: 1 =drank alcohol during pregnancy, 0 otherwise
   - *smoking*: 1 =smoked during pregnancy, 0 otherwise

   A first-order multinomial logistic regression is carried out for this study, and the R output is shown below.

```
Call:
multinom(formula = preg ~ nutrition + less20 + greater30 + alcohol +
    smoking)

Coefficients:
             (Intercept)   nutrition   less20 greater30  alcohol  smoking
preterm         5.475147 -0.06541919 2.957028  2.059662 2.042900 2.452362
intermediate    3.958370 -0.04644903 2.913475  1.887550 1.067001 2.230492

Std. Errors:
             (Intercept)  nutrition    less20 greater30   alcohol   smoking
preterm         2.271677 0.01823916 0.9644921 0.8947727 0.7097461 0.7315106
intermediate    1.941063 0.01488581 0.8575544 0.8088255 0.6495262 0.6681955
```

   (a) Write down the estimated multinomial logistic regression models associated with this analysis.

   (b) Calculate the Wald test statistics associated with the predictor *alcohol*, and find the corresponding p-value. What are the conclusions in context at significance level $\alpha = 0.05$? You do not need to apply the Bonferroni method here.

   (c) Calculate the 95% confidence intervals associated with the predictor *alcohol*, and interpret these intervals in context, in terms of relative risk of having a pregnancy that is pre-term, intermediate, or full term. You do not need to apply the Bonferroni method here.

3. **This question is optional** (No R required) Recall that a probability distribution belongs to the exponential family if its distribution function takes the following form:

$$f(y;\theta) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where θ is the canonical parameter and ϕ is the dispersion parameter. Show that the Poisson distribution belongs to the exponential family. Also, since we know that for exponential families, $\mathrm{E}\{y\} = b'(\theta)$ and $\mathrm{Var}\{y\} = a(\phi)b''(\theta)$, obtain the expected value and variance for the Poisson distribution.

As a reminder, the probability mass function of a Poisson distribution is

$$f(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

for $y \in \{0, 1, 2, 3, \cdots\}$.