Diet and Colon Cancer

In this example, you will study data obtained from a state-of-the-art colon carcinogenesis scientific experiment. The objective is to estimate the probability of programmed cell death (apoptosis) for cells at different ages to study the effect of various diets on the progression of colon cancer. An insufficient amount of apoptosis results in uncontrolled cell proliferation, the leading cause of cancer.

In the experiment, a number of 12 rats were fed four different diets (corn oil or fish oil diet and with or without butyrate supplementation), exposed to a carcinogen that induces colon cancer. The colon crypts at different locations of the colon tissue for the 12 rats were examined. A crypt is typically 20-30 cells deep, where the cell location is a measure of its age. To simplify our interpretation, we define the relative cell position as $t = 0$ at the bottom of each crypt (less mature cells) and $t = 1$ at the top of each crypt (more mature cells). In these data, the colon has been sampled at about 20 crypt locations per rat ($M = 20$) with about 25 cells per crypt ($N \approx 25$). There are approximately 10% apoptotic cells of the total cells examined - a rare event setting. Further details about the experiment and data are provided in Baladandayuthapani et al. (2008).

The data consists of the following variables:

- *Response*: Whether the cell died (=1) or not (=0)
- *Rat ID*: a number between 1 and 12 specifying identity number of the rats
- *Diet ID*: a number between 1 and 4 specifying identity the type of diet ()
- *Age*: Age of the cell where $t = 0$ at the bottom of each crypt (less mature cells) and $t = 1$ at the top of each crypt (more mature cells)
- *Crypt*: The location of the crypt in the colon

Getting the Data: Go to the T-square class site and in the Resources section, you will find the dataset for this practice exam called *diet_and_cancer.csv*. Once you have saved the data file in the working directory, read the data in R using the command

```
data = read.csv("diet_and_cancer.csv",header=TRUE)
```

In this study, you will investigate whether apoptosis or cell death is associated to diet while controlling for other cell characteristics and variations due to observations in different rats, for cells at different age and from crypts in different locations of the colon. If the association is statistically significant, it is important to assess which diet leads to a lower probability of cell death, which could be an indicator that the diet indeed helps in preventing or slowing down cancer.

Some steps you may take in analyzing these data:

- Rat id and diet id are categorical variables and you should be careful in incorporating them in the model as such; for this you will need to create dummy variables similarly as we did in multiple regression analysis.
- Use exploratory techniques, graphics and summaries, to analyze the data before applying generalized linear regression.
- Along with estimation, you should perform variable selection. You may consider also interaction between predictors.
- You will need to finally assess the goodness-of-fit.
- The most important part is interpretation and findings!

Note: Since the response is binary but rare-event, in the epidemiology literature is common to use the log link function instead of the logistic function. Compare the results of your model based on the two link functions.

Deliverable: Write a report for the analysis of this study. Be organized. Write in plain but correct English.