STA 303H1S / STA 1002HS: Logistic Regression 2 Practice Problems

1. (a) To confirm the appropriateness of the logistic regression model $\text{logit}(\pi) = \beta_0 + \beta_1 x$, it is sometimes useful to fit $\text{logit}(\pi) = \beta_0 + \beta_1 x + \beta_2 x^2$ and test whether β_2 is zero.
   i. Does the reliability of this test require large values of m_i, the number of observations for each value of x?
   ii. Is the test more relevant when the m_i are small?

   (b) On the practice problems website there is R code and output for the Krunnit Islands example, including the square of the log of area as a predictor variable. Evaluate, using as many statistics as you can, whether or not this model provides a superior fit to the data than the model examined in lecture.

   (c) Logistic regression analysis assumes that, after the effects of explanatory variables have been accounted for, the responses are independent of each other. Why might this assumption be violated for the Krunnit Islands examples?

2. Consider logistic regression with one explanatory variable. In logistic regression with a binomial response, y_i is a count of the number of events in m_i trials for the ith value of the explanatory variable. If the response is binary, y_i is 1 is the event happens for the ith observation and 0 otherwise. In both cases, logistic regression models the logit of the probability of the event occurring as $\beta_0 + \beta_1 x_i$. Each observation in the binomial response case can be expanded into m_i binary responses, y_i of which are 1 and $m_i - y_i$ of which are 0. Explain why the maximum likelihood estimates of β_0 and β_1 are the same for the binomial response model and the binary response model fit to the expanded data.

3. (Adapted from questions 14.11 and 14.17 of Kutner *et al.*)
   A carefully controlled experiment was conducted to study the effect of the size of the deposit level on the likelihood that a returnable one-litre soft-drink bottle will be returned. A bottle return was scored 1, and no return was scored 0. The data for this question are the number of bottles returned out of 500 sold at each of six deposit levels (in cents). Relevant R output and plots are on the practice problems website.

   (a) Consider the plot of the estimated response proportions ($\hat{\pi}_S$) against deposit level. Does the plot support the appropriateness of a logistic response function that is a linear function of deposit level?

   (b) Consider the plot of the logit of the estimated response proportions ($\hat{\pi}_S$) against deposit level. Does the plot support the appropriateness of a logistic response function that is a linear function of deposit level?

   (c) What is the fitted model?

(d) Consider the plot of the estimated response proportions ($\hat{\pi}_S$) against deposit level with the proportions estimated from the model ($\hat{\pi}_M$) superimposed. Does the fitted logistic response function appear to fit well?

(e) Obtain $\exp(\hat{\beta}_1)$ and interpret this number.

(f) What is the estimated probability that a bottle will be returned when the deposit is 15 cents?

(g) Estimate the amount of deposit for which 75% of the bottles are expected to be returned.

(h) Obtain an approximate 95% confidence interval for β_1. Convert this confidence interval into one for the odds ratio. Interpret this latter interval.

(i) Conduct a Wald test to determine whether deposit level is related to the probability that a bottle is returned.

(j) Conduct a likelihood ratio test to determine whether deposit level is related to the probability that a bottle is returned.

(k) The Deviance residuals are printed. What do you conclude from them?

(l) Suppose we treated deposit level as a categorical variable instead. Conduct a Deviance Goodness-of-Fit test to see if the linear model is appropriate.