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

1. (a) i. Maybe. The Wald test and the likelihood ratio test both can be used if either the sample size is large or the m_i are large.

   ii. The test is useful as an informal device for assessing the model. It may be more useful when the m_i are small, since few alternatives are available for model checking in this case.

   (b) All of the following statistics give evidence that the square of the log of area is not needed in the model:
   - The Wald test with null hypothesis that the coefficient of the square of the log of area is 0 has p-value 0.7736.
   - The likelihood ratio test with null hypothesis that the coefficient of the square of the log of area is 0 has test statistic 0.082 with p-value 0.7742. So the data are consistent with a 0 coefficient.
   - AIC for the model with the square of the log of area (77.311) is greater than AIC for the model without it (75.394).
   - SC for the model with the square of the log of area (79.98) is greater than SC for the model without it (77.17).

   (c) Neighbouring islands are likely to be more similar in the types of species present and in the likelihood of extinction for each species than islands farther apart. Another possibility is that extinctions of birds can occur in another part of the world and simultaneously affect all or several of the Krunnit Islands populations.

2. In the binomial response case, the log-likelihood function is

$$\log(L) = \sum_{i=1}^{n} \left(\log \begin{pmatrix} m_i \\ y_i \end{pmatrix} + y_i \log(\pi_i) + (m_i - y_i) \log(1 - \pi_i) \right)$$

where $\pi_i = \exp(\beta_0 + \beta_1 x)/(1 + \exp(\beta_0 + \beta_1 x))$, and the estimates of β_k, $k = 0, 1$, are found by solving the equations

$$\frac{\partial \log(L)}{\partial \beta_k} = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i} \frac{\partial \pi_i}{\partial \beta_k} + \frac{m_i - y_i}{1 - \pi_i} \left(-\frac{\partial \pi_i}{\partial \beta_k} \right) \right) = 0$$

In the binary response case, all m_i's are 1's, but otherwise the equations are the same, except for the interpretation of n and the y_i's.

Expanding the binomial case, replace each y_i with m_i observations y'_{ij} where $j = 1, \ldots, m_i$ where y_i of these are 1's and $m_i - y_i$ of these are 0's and the equations to be solved become

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(\frac{y'_{ij}}{\pi_i} \frac{\partial \pi_i}{\partial \beta_k} + \frac{1 - y'_{ij}}{1 - \pi_i} \left(-\frac{\partial \pi_i}{\partial \beta_k} \right) \right) = 0$$

and the solution corresponds to the binary case.

3. (a) It seems reasonable that an S-shaped logit function would fit this plot well.
   (b) A linear model seems appropriate from this plot.
   (c) $\text{logit}(\hat{\pi}) = -2.0763 + 0.1358\,\text{deposit}$
   (d) Looks pretty good.
   (e) $\exp(\hat{\beta}_1) = 1.145$. An increase in deposit level of 1 cent is associated with a 14.5% increase in the odds that a bottle will be returned.
   (f) $\hat{\pi} = \frac{\exp(-2.0763+0.1358(15))}{1+\exp(-2.0763+0.1358(15))} = 0.49$
   (g) Solving $\log\left(\frac{0.75}{1-0.75}\right) = -2.0763 + 0.1358\,\text{deposit}$ gives an estimate of a deposit of 23.4 cents for 75% of bottles to be returned.
   (h) The 0.025 quantile from a standard normal distribution is 1.96.
   An approximate 95% confidence interval for β_1 is $0.1358 \pm 1.96(0.00477) = (0.126, 0.145)$. Exponentiate this to get the approximate 95% confidence interval for the odds ratio which gives (1.135, 1.156). Thus for each 1 cent increase in deposit level, we estimate that the odds that a bottle will be returned increase by a value in the range 13.5% to 15.6%. (Note that this confidence interval is given in the R output.)
   (i) The p-value from R is < 0.0001 so there is strong evidence that deposit level is related to the probability that a bottle is returned.
   (j) The appropriate test here is the likelihood ratio test for the global null hypothesis. From R this has a p-value of < 0.0001 so there is strong evidence that deposit level is related to the probability that a bottle is returned. (Note that you should be able to get the test statistic (1095.99) from other numbers available from R.)
   (k) Look for outliers. The deviance residual for a deposit of 20 cents is almost 3, so this deposit level is not well fit by the model.
   (l) The test statistic for the Deviance Goodness-of-Fit test is 12.181. Under the null hypothesis that the linear model is appropriate for the log-odds, this is an observation from a chi-square distribution with 4 degrees of freedom. The estimated p-value from the chi-square table is between 0.01 and 0.025. So there is evidence that the linear function is not appropriate.