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

SOLUTIONS

1. (a) There were no females over 50. Any comparisons for older people must be based on the assumption that the model still holds and this cannot be verified with these data. Extrapolation is a problem for logistic regression, just as it is for linear regression.

   (b) Males and females might have different tasks and survival could be associated with task.

   (c) i. $3.2 - 0.078 age - 1.6 I_{male}$

   ii. Same equation but with all coefficients negative.

   (d) Estimated probabilities for women are

   $$\hat{\pi} = \frac{e^{1.6-0.078 age+1.6}}{1+e^{1.6-.078 age+1.6}}$$

   and for men are

   $$\hat{\pi} = \frac{e^{1.6-0.078 age}}{1+e^{1.6-.078 age}}$$

   So estimated survival probabilities for males are 0.41 at age 25 and 0.091 at age 50. For females, the estimates are 0.78 at age 25 and 0.33 at age 50.

   (e) Set the estimated log-odds to zero and solve for age. For females, the age of 50% survival is 41.0 years; for males it is 20.5 years.

   (f) Since females are coded as 1 and males as -1, females versus males corresponds to a change of 2 in the explanatory variable `sex`. So the odds ratio for females versus males (of the same age) is found by exponentiating 2 times the estimated coefficient for `sex`.

   (g) The test statistic is $13.326 - 10.127 = 3.199$. Under the null hypothesis that the reduced model adequately fits the data, this is an observation from a chi-square distribution with 1 degree of freedom (since there is 1 less parameter in the reduced model), giving an estimated p-value between 0.05 and 0.1. So there is weak evidence that the reduced model is not adequate; that is, there is weak evidence that 30 is not the age at which the probability of survival for females is 0.5.

   (h) i. The deviance for the model with the interaction term is 47.346 and for the model without the interaction term the deviance is 51.256. The LRT with null hypothesis that the coefficient of the interaction term is zero has test statistic $51.256 - 47.346 = 3.91$. Under the null hypothesis this is an observation from a chi-square distribution with 1 degree of freedom. From tables, we can estimate the p-value as $0.025 < p < 0.05$. So there is moderate evidence that the coefficient of the interaction term is not 0.

   ii. (1) The estimated odds ratio is $e^{-.0325(10)-0.1616(10)} = 0.14$.

   (2) The estimated odds ratio is $e^{6.9267-0.1616 age}$. Note that it differs with the age.

   iii. No. If there is no interaction term, $\exp(\beta_1)$ is the odds ratio for a person that is 1 year older than another person, but in the presence of an interaction term, the odds ratio must be adjusted for gender. With an interaction, $\exp(\beta_1)$ is the odds ratio for a 1 year change in age for males only.

(i)

Explanatory variables in model	AIC	Rank by AIC	SC	Rank by SC
age, sex, age*sex	55.346	1	62.573	1
age, sex, age*sex, age^2	55.830	2	64.863	3
age, sex	57.256	3	62.676	2
age, sex, age*sex, age^2, age$^2 *$ sex	57.361	4	68.201	4

Both AIC and SC choose the model with age, sex and their interaction as the model that best fits the data. This is not surprising, considering that all of these have coefficients that are statistically significantly different from 0 (using the likelihood ratio test for the interaction term). Both AIC and SC least favour the most complicated model. The other models are ranked differently, reflecting that SC penalizes more complicated models more severely than AIC. As a rule-of-thumb, a difference of 2 or less in AIC indicates no real difference in the fit of the model (a difference in AIC greater than 10 is considered an important difference), so there is no overwhelming evidence from AIC in favour of any of these models. A parsimonious approach would be to select the simplest model.

2. (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 $544.736 - 544.654$ which is very small. So the data are consistent with a 0 coefficient.
   - AIC for the model with the square of the log of area (550.654) is greater than AIC for the model without it (548.736).
   - SC for the model with the square of the log of area (564.000) is greater than SC for the model without it (557.634).

   (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.

3. $\log\left(\frac{Y_i}{1-Y_i}\right)$ is undefined if $Y_i = 1$ or 0.

4. 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.

5. (a) Maximum likelihood estimates are: $\hat{\beta}_0 = -4.7393$, $\hat{\beta}_1 = 0.0677$, and $\hat{\beta}_2 = 0.5986$. The fitted response function is $\text{logit}(\hat{\pi}) = -4.74 + 0.068\,\text{income} + 0.60\,\text{age}$.

   (b) $\exp(\hat{\beta}_1) = 1.07$ so an increase of \$1000 in income is associated with a 7% increase in the odds of purchasing a new car in the next year.
   $\exp(\hat{\beta}_2) = 1.82$ so an increase of 1 year in the age of the oldest family automobile is associated with an 82% increase in the odds of purchasing a new car in the next year.

   (c) $\hat{\pi} = \frac{\exp(-4.74+0.068(50)+0.60(3))}{1+\exp(-4.74+0.068(50)+0.60(3))} = 0.61$

   (d) The 0.005 quantile from a standard normal distribution is 2.576. An approximate 99% confidence interval for the coefficient of income is $0.0677 \pm 2.576(0.0281) = (-0.0047,\ 0.14)$. An approximate 99% confidence interval for the odds ratio for families whose incomes differ by \$20 thousand is $(e^{-0.0047(20)},\ e^{0.14(20)}) = (0.91,\ 16.5)$.

   (e) $H_0:\ \beta_2 = 0$ versus $H_a:\ \beta_2 \neq 0$.
   The test statistic is $(0.5986/0.3901)^2 = 2.35$. Under the null hypothesis, this is approximately an observation from a chi-square distribution with 1 degree of freedom. SAS gives the p-value as 0.1249. From a chi-square table we can say that $0.1 < p < 0.9$. We can get a more accurate estimate of the p-value by using the fact that $0.5986/0.3901 = 1.53$ is approximately an observation from a standard normal distribution. The p-value is then $2(1 - 0.9370) = 0.126$. So the data are consistent with the coefficient of age being 0.

   (f) The models to be compared are $\text{logit}(\pi) = \beta_0 + \beta_1\,\text{income} + \beta_2\,\text{age}$ and $\text{logit}(\pi) = \beta_0 + \beta_1\,\text{income}$. The likelihood ratio test statistic is $39.305 - 36.690 = 2.615$. The estimated p-value from the chi-square table with 1 degree of freedom is $0.1 < p < 0.9$. The conclusion is consistent with the Wald test.

   (g) The test statistic is $36.690 - 34.253 = 2.437$. Under the null hypothesis that the coefficients of all the second-order terms are zero, this is approximately an observation from a chi-square distribution with 3 degrees of freedom. From the chi-square table, the p-value is between 0.1 and 0.9 we conclude that no significant contribution to the model is being made by the 3 second-order terms.

6. (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 SAS output.)

   (i) The p-value from SAS 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 SAS 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 SAS.)

   (k) Look for outliers. Both the deviance and Pearson residuals 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 $3062.872 - 3050.690 = 12.182$. 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.

7. (a) For the Donner Party example, the cut-off probabilities of 0.50 and 0.52 correctly classify the status of 73.3% of people. Associated with these cut-offs are fairly low false positive (16.7%) and false negative (30.3%) rates, so 0.50 is a reasonable choice for a cut-off probability balancing the criteria equally.

   For the Krunnit Islands example, the cut-off probabilities of 0.34, 0.36, 0.38 and 0.42 all classify 82.9% of species correctly. The cut-off probability of 0.42 classifies all of the non-extinct species correctly and all of the extinct species incorrectly; the other cut-offs classify 3 (of 108) of the extinct species correctly and 521 (of 524) of the non-extinct species correctly and give the lowest false positive rates. Low false negative rates result in a small overall percent correct.

   (b) 0.50 is a reasonable cut-off probability for the Donner Party example but not for the Krunnit Islands example. This is because we have observed close to an equal number of events/non-events for the Donner Party example but not for the Krunnit Islands example. Since we observed fewer non-events for the Krunnit Islands example, we achieve better classification rates by choosing a cut-off probability that is smaller.

(c) These classification rates are calculated from the data that were used to fit the model and thus will be overly optimistic. It would be better to fit the model on a training data set, and estimate the classification rates from a test data set.

8. (a) The SAS output warns us that “Complete separation of data points detected.” and “The maximum likelihood estimate does not exist.”

(b) Separation occurs because a straight line can be drawn in the bottom versus diagonal plot that completely separates the banknotes into genuine and counterfeit. So a linear function of bottom and diagonal perfectly classifies the banknotes.