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

1. This question relates to the Donner Party example from lecture. The SAS output you need is on the SAS examples from lecture web page; additional SAS output is also given on the practice problems web page for part (g).

   (a) Why should one be reluctant to draw conclusions about the ratio of male and female odds of survival for the Donner Party members over 50?

   (b) What confounding variables might be present that could explain the significance of the sex indicator variable?

   (c) From the Donner Party data, the log odds of survival were estimated to be $1.6 - 0.078 age + 1.6 I_{female}$, based on a binary response that takes the value 1 if an individual survived and I_{female} is an indicator variable that takes the value 1 for females.

      i. What would be the estimated equation for the log-odds of survival if the indicator variable for sex were 1 for males and 0 for females?

      ii. What would be the estimated equation for the log-odds of perishing if the binary response were 1 for a person who perished and 0 for a person who survived?

   (d) What are the estimated probabilities of survival for men and women of ages 25 and 50?

   (e) What is the age at which the estimated probability of survival is 50% for women and for men?

   (f) The odds ratio for a unit change in an explanatory variable, holding other explanatory variables constant, is typically estimated by exponentiating the estimated coefficient of the explanatory variable. Consider the second model fit on the lecture examples web page for the Donner Party example. In this model, we modeled the logit of probability of survival as a linear function of age and sex, where sex was coded as 1 for females and -1 for males. In the corresponding output, SAS gives 4.940 as the estimated odds ratio for "`sex FEMALE vs MALE`". Explain how this is calculated from other numbers in the SAS output.

   (g) Consider the Donner Party females (only) and the logistic regression model $\beta_0 + \beta_1 age$ for the logit of survival probability. If A represents the age at which the probability of survival is 0.5, then $\beta_0 + \beta_1 A = 0$. This implies that $\beta_0 = -\beta_1 A$ (why?). The hypothesis that $A = 30$ years may be tested by the drop-in-deviance test with the following reduced and full models for the logit:
   Reduced: $-\beta_1 30 + \beta_1 age = 0 + \beta_1(age - 30)$
   Full: $\beta_0 + \beta_1 age$
   To fit the reduced model, one must subtract 30 from the ages and drop the intercept term. The drop-in-deviance test statistic is computed in the usual way. Carry out the test that $A = 30$ for the Donner Party females.

   (h) Consider the model including age, sex, and their interaction.

      i. Use a Likelihood Ratio Test to test whether the coefficient for the age-sex interaction is statistically significantly different from 0.

      ii. Find the estimated odds ratio for: (1) a female that is 10 years older than another female, and (2) females versus males of the same age.

iii. Does $\exp(\beta_1)$ have the same meaning here as for a model containing no interaction term?

(i) In lecture we looked at several models for the Donner Party data, including age, sex, their interaction, a quadratic term in age, and the interaction of sex and the quadratic term. Rank the various models according to AIC and SC. What do you conclude?

2. (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 SAS 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?

3. (Question 14.2 in Kutner *et al.*)
   Consider logistic regression with a binary response. Since the logit transformation linearizes the logistic response function, why can't this transformation be used on the individual responses Y_i and a linear response function then fitted?

4. 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.

5. (Adapted from questions 14.13 and 14.19 of Kutner *et al.*)
   A marketing research firm was engaged by an automobile manufacturer to conduct a pilot study to examine the feasibility of using logistic regression for ascertaining the likelihood that a family will purchase a new car during the next year. A random sample of 33 suburban families was selected. Data on annual family income (`income`, in thousand dollars) and the current age of the oldest family automobile (`age`, in years) were obtained. A follow-up interview conducted 12 months later was used to determine whether the family actually purchased a new car (`purchase`= 1) or did not purchase a new car (`purchase`= 0) during the year. A logistic regression model with two predictor variables in first-order terms (*i.e.*, no polynomial terms) is assumed to be appropriate. Assume also that large-sample inferences are applicable. Relevant SAS output is on the practice problems web site.

   (a) What are the maximum likelihood estimates of β_0, β_1, and β_2? State the fitted response function.

(b) Obtain $\exp(\hat{\beta}_1)$ and $\exp(\hat{\beta}_2)$ and interpret these numbers.

(c) What is the estimated probability that a family with annual income of $50 thousand and an oldest car of 3 years will purchase a new car next year?

(d) Find a 99% confidence interval for the family income odds ratio for families whose incomes differ by 20 thousand dollars.

(e) Use the Wald test to determine whether `age` can be dropped from the model. State the null and alternative hypotheses. Verify the p-value given by SAS by estimating it as precisely as possible from an appropriate statistical table.

(f) Use the likelihood ratio test to determine whether `age` can be dropped from the model. What models are being compared in this test? How does the result compare to that obtained for the Wald test in the previous part?

(g) Use the likelihood ratio test to determine whether the following three second-order terms, the square of family income, the square of age of oldest automobile, and the two-factor interaction effect between annual family income and age of oldest automobile, should be added simultaneously to the model containing family income and age of oldest automobile as first-order terms.

6. (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 SAS 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 and Pearson residuals are printed. What do you conclude from them?

(l) SAS output is also given for a logistic regression model treating deposit level as a categorical variable. Use this output to conduct a Deviance Goodness-of-Fit test.

7. Logistic regression is one possible method to find a combination of explanatory variables to best classify observations into two groups. An observation is classified as having estimated response of 1 if the estimated probability of 1 from the logistic regression model is greater than a cut-off probability; else it is classified as having an estimated response of 0. The `ctable` option in `proc logistic` in SAS prints a table of cut-off probabilities and the resulting classifications. The output it produces includes:
- `Prob level`: a sequence of proposed cut-off probabilities in increments of 0.02
- `Correct Event`: the number of events (observations for which $Y = 1$) that are classified as events using the given cut-off probability
- `Correct Non-Event`: the number of non-events (observations for which $Y = 0$) that are classified as non-events using the given cut-off probability
- `Incorrect Event`: the number of non-events that are classified as events using the given cut-off probability
- `Incorrect Non-Event`: the number of events that are classified as non-events using the given cut-off probability
- `Correct`: the percentage of observations that are classified correctly using the given cut-off probability
- `False POS`: the percentage of observations classified as events that are incorrectly classified
- `False NEG`: the percentage of observations classified as non-events that are incorrectly classified
- You can ignore the sensitivity (percentage of events correctly classified) and specificity (percentage of non-events correctly classified).

On the practice problems website, output including Classification Tables is given for both the Donner Party and Krunnit Islands examples.

(a) For each example, choose the optimal cut-off probability considering the percentage correct and the percentages of false positives and false negatives.

(b) Without any additional information, you might guess that 0.5 is a reasonable choice for the cut-off probability. Are your answers to part (a) close to 0.5? Why or why not?

(c) Why should you not trust the percentage of correct classifications?

8. (Adapted from Question 8.6 of Sheather.)
A statistician at the University of Berne was asked by local authorities to analyze data on Swiss Bank notes. In particular, the statistician was asked to develop a model to predict whether a particular banknote is counterfeit ($y = 0$) or genuine ($y = 1$) based on the following physical measurements (in millimetres) of 100 genuine and 100 counterfeit Swiss Bank notes:
length = length of the banknote
left = length of the left edge of the banknote
right = length of the right edge of the banknote

top = distance from the image to the top edge
bottom = distance from the image to the bottom edge
diagonal = length of the diagonal

The data and some SAS code and output are on the practice problems website.

(a) Fit a logistic regression model using just the last two predictor variables listed (bottom and diagonal). What is unusual about the SAS output?

(b) Look at the plot of bottom versus diagonal, coded by whether or not the banknote is genuine. How is what you see in the plot related to your answer to part (a)?