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

1. This question relates to the Donner Party example.

   (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 $1.6 - 0.078age + 1.6I_{female}$, 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 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. The estimated odds ratio for "`sex FEMALE vs MALE`" is 4.94. Explain how this is calculated from other numbers in the R 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 $\beta_0 + \beta_1 A = 0$ (why?). This implies that $\beta_0 = -\beta_1 A$. 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) 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. (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?

3. (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 R 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) Write R code (without using `fit` or `glm` or anything like that) to 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 in R.

(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.

4. (Adapted from Question 8.6 of Sheather.)
A statistician at the University of Bern 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

   (a) We fit a logistic regression model using just the last two predictor variables listed (bottom and diagonal). What is unusual about the R 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)?