Homework # 4

Due Wednesday 3/21

Turn in R code where applicable.

1. *Complete and Quasi-complete Separation:* Read Section 6.5.1 of the textbook. Consider the data $y = (0, 0, 0, 0, 1, 1, 1, 1)$ and $x_1 = (1, 2, 3, 3, 5, 6, 10, 11)$.

   (a) Fit a logistic model with response y and predictor x_1. Explain the warning message from R. Describe the model fit and coefficient estimates and standard errors. Is there evidence of complete or quasi-complete separation?

   (b) Now consider the same logistic model but with $x_1 = (1, 2, 3, 3, 3, 6, 10, 11)$. Explain the warning message from R. Describe the model fit and coefficient estimates and standard errors. Is there evidence of complete or quasi-complete separation?

2. *Grouped vs. Ungrouped Data:* A study has n_i independent binary observations $\{y_{i1}, \ldots, y_{in_i}\}$ when $X = x_i$, $i = 1, \ldots, N$ with $n = \sum_{i=1}^{N} n_i$. Consider the model $logit(\pi_i) = \beta_0 + \beta_1 x_i$, where $\pi_i = P(Y_{ij} = 1)$.

   (a) Show that the kernel of the likelihood function is the same treating the data as n Bernoulli observations or N binomial observations.

   (b) For the saturated model, explain why the likelihood function is different for these two data forms. Hence, the deviance reported by R depends on the form of the data entry.

   (c) Explain why the difference between deviances for the two unsaturated models does not depend on the form of the data entry.

   (d) Suppose that each $n_i = 1$. Show that the deviance depends on $\hat{\pi}_i$ but not y_i. Hence it is not useful for checking model fit.

3. *Logistic Regression.* The Donner Party was the most famous tragedy in the history of the westward migration in the United States. In the winter of 1846-47, about ninety wagon train emigrants were unable to cross the Sierra Nevada Mountains of California before winter, and almost one-half starved to death. This data, `donner.txt` posted on Blackboard, includes information about each of the members of the party: survival status (1 if survived, 0 if died), sex (1 if male, 0 if female), age, and status (family, single or hired).

   (a) Fit a logistic model with response *survival* and predictor *age.*

   (b) State and interpret $\hat{\beta}_{age}$ and $e^{\hat{\beta}_{age}}$.

   (c) What are the results from the Wald test of the relation between *survival* and *age*? State the null hypothesis, alternative hypothesis, z-statistic, p-value and conclusion. Assume $\alpha = .05$.

   (d) State the 95% confidence intervals for $\hat{\beta}_{age}$ and $e^{\hat{\beta}_{age}}$. Do these indicate that the effect of age on survival is statistically significant? Explain.

   (e) Fit a logistic model with response *survival* and predictors *age*, *age*^{2}, *sex* and *status.*

   (f) Formally assess whether the model in part (e) is better than the original model from part (a). State the null hypothesis, alternative hypothesis, likelihood ratio test statistic (using the deviance), rejection region and conclusion. Assume $\alpha = .05$.

(g) State and interpret $\hat{\beta}_{sex}$ and $e^{\hat{\beta}_{sex}}$.

4. *Prediction in Logistic Regression.* This problem uses the Donner Party data from Question 3.

   (a) Using the model from part (e) of Question 3, obtain the fitted values for the log odds of survival. Use the `summary` command to show descriptive statistics of these predictions.

   (b) Use

   $$\log\left(\tfrac{\pi(x_i)}{1-\pi(x_i)}\right) = x_i^T\beta$$

   to show that

   $$\theta(x_i) = \tfrac{exp(x_i^T\beta)}{1+exp(x_i^T\beta)}$$

   where $\pi(x_i) = P(survival = 1|X = x_i)$, x_i is the vector of predictors for observation i and β is the vector of parameters.

   (c) Using the model from part (e) of Question 3, obtain the fitted values for the predicted probability of survival. Use the `summary` command to show descriptive statistics of these predictions.

   (d) In order to get binary predictions, a cutoff point for the predicted probability of survival must be chosen: below which the outcome of 0 is predicted and above which the outcome of 1 is predicted. In this case, the outcome (denoted y) is the variable *survival.* Here you will compare two cutoffs:

      i. Use 0.5 as a cutoff. This corresponds to the prediction rule:

      $$\hat{y}_i = \begin{cases} 1 \text{ if } \hat{\pi}(x_i) > .5 \\ 0 \text{ otherwise} \end{cases}$$

      Provide a cross-tabulation of y and $\hat{y}$ using the `table` command in R. How many of the observations have incorrect predictions (i.e. the predicted value is not equal to the observed value)?

      ii. Use the proportion of survivors from the data as a cutoff. This corresponds to the prediction rule:

      $$\hat{y}_i = \begin{cases} 1 \text{ if } \hat{\pi}(x_i) > \frac{\sum_{i=1}^n y_i}{n} \\ 0 \text{ otherwise} \end{cases}$$

      Provide a cross-tabulation of y and $\hat{y}$ using the `table` command in R. How many of the observations have incorrect predictions (i.e. the predicted value is not equal to the observed value)?

      iii. Comparing the proportion of incorrect predictions is one way determine the cutoff. For the two rules used above, calculate the proportion of incorrect predictions. Based on the proportion of incorrect predictions, which of the two cutoffs would you choose to use?

      iv. Calculate the false positive rate (FPR) and false negative rate (FNR) for each of the two cutoffs, where:

      $$FPR = P(\hat{y} = 1|y = 0)$$
      $$FNR = P(\hat{y} = 0|y = 1)$$

Note that probabilities can be estimated by proportions observed in the data in this case.

v. Based on the false positive and false positive rates, which of the two cutoffs would you choose to use to define binary predictions? Note that a smaller false positive rate and a smaller false negative rate is preferred.

vi. Plot the ROC curve for the model from part (e) of Question 3. Describe the predictive power of the model based on the ROC plot. Mark the points on the curve that correspond to the chosen cutoff π_0 in part (i) and (ii) (i.e. .5 and $\overline{y}$). *You can write your own function or use an R package for ROC curves.*