Statistical Computing - Exercises 09 - Logistic Regression

We will implement Fisher scoring for logistic regression, and apply it to the 2003 NFL field goal data.

Model for Binary Responses

We have binary data (0s and 1s) $y_1, \dots, y_n$, and continuous covariates x_{ij} (jth covariate for the ith observation). We model the responses in the following way:

$$P(Y_i = 1) = p_i$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=0}^{J} b_j x_{ij}$$

where $x_{i0} = 1$ if we have an intercept in the model.

It works similarly to a regular regression, except we model the log of the odds of a success as a linear function of the covaraites.

Exercises

1. Show that the probability mass function can be written like this:

$$f(y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

2. Show that

$$p_i = \frac{\exp\left(\sum_{j=0}^{J} b_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{J} b_j x_{ij}\right)}$$

3. Derive the negative loglikelihood function for all observations (this should involve a sum over i).
4. Derive the gradient of the negative loglikelihood with respect to $b_0, \dots, b_J$.
5. Derive the Fisher information.
6. Read the 2003 NFL field goal data `datasets/fieldgoal2003.csv` into R. Each row represents an attempt to kick a field goal in an NFL game in the 2003 regular season. The first column is the number of yards from the goal, the second column indicates whether the attempt was successful (1 = success, 0 = failure), and the third column is the week of the regular season (1-17).

   Make an informative plot of the data. You may want to experiment with aggregating data to specific values of the yards, and incorporating information about the number of attempts from each value of yards.
7. Code up the Fisher scoring algorithm in R, and apply it to the 2003 NFL field goal data. Use yards and week as the two covariates, and include an intercept. Print the values of the regression coefficients and the loglikelihood function at each iteration. Show that it converges from different choices of starting values.
8. How do you interpret your results? Are the regression coefficients surprising?
9. Fit the same model using the `glm` function in R to verify that you got the right answer. If you haven't used the `glm` function before, you may need to read its documentation. You will need to use argument `family = binomial`.