Generalized Linear Models

Problem 1 (25 points: 5 points each question): Building and analyzing the logistic regression model

For the problem below, build the logistic regression model (fit.all) using all the predictors and answer the following questions by including the corresponding R code and showing all the required mathematical derivations used to answer these questions:

1. Let X_h be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient. Build a single predictor logistic regression model (*fit.single*) using X_h as the predictor. Write the equations relating the dependent variable (Response) to the explanatory variable in terms of:
    a. Probabilities: $Prob(Y = Yes \mid X_h = x)$
    b. Odds: $Prob(Y = Yes)$
    c. Logit

2. Write the estimated equation for the *fit.all* model in all three formats (if the number of predictors is more than four, then include only those four predictors whose absolute value estimates are the highest):
    a. The logit as a function of the predictors.
    b. The odds as a function of the predictors.
    c. The probability as a function of the predictors

3. Let X_h be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient in the *fit.all*. Compute the odds ratio that estimated a single unit increase in X_h, holding the other predictors constant. For example, if $X_h = 1\ then$:
$$\frac{odds(X_1 + 1, X_2, \dots, X_q)}{odds(X_1, X_2, \dots, X_q)} =$$
Provide the interpretation for this regression coefficient. If it were a linear regression model, how would the interpretation change for a single unit increase in X_h.

4. Build a reduced logistic regression model (*fit.reduced*) using only the predictors that are statistically significant. Assess if the reduced model is equivalent to the full model. Justify your answer.

5. Compute the dispersion of your model and run the dispersion diagnostic test. If the constructed model is overdispersed, then discuss the ways to deal with the issue.

Competitive Auctions on eBay.com. The file eBayAuctions.xls contains information on 1972 auctions transacted on eBay.com during May–June 2004. The goal is to use these data to build a model that will distinguish competitive auctions from noncompetitive ones. A competitive auction is defined as an auction with at least two bids placed on the item being auctioned. The data include variables that describe the item (auction category), the seller (his or her eBay rating), and the auction terms that the seller selected (auction duration, opening price, currency, day of week of auction close). In addition, we have the price at which the auction closed. The goal is to predict whether or not the auction will be competitive.

Data Preprocessing. Create dummy variables for the categorical predictors. These include Category (18 categories), Currency (USD, GBP, euro), EndDay (Monday–Sunday), and Duration (1, 3, 5, 7, or 10 days). Split the data into training and validation datasets using a 60% : 40% ratio.

a. Create pivot tables for the average of the binary dependent variable (Competitive?) as a function of the various categorical variables (use the original variables, not the dummies). Use the information in the tables to reduce the number of dummies that will be used in the model. For example, categories that appear most similar with respect to the distribution of competitive auctions could be combined.