Assignment #7: Problem Set for Logistic Regression (50 points)

This assignment will be made available in both pdf and Microsoft docx format. Answers should be typed into the docx file, saved, and converted into pdf format for submission into Blackboard. **Color your answers in green so that they can be easily distinguished from the questions themselves.**

All of these computations are covered in examples in the assigned reading, and hopefully in the notes that you have been taking. If you need to refresh your memory, then begin by looking at Chapter 12 in *Regression Analysis By Example* and Chapter 1 in *Applied Regression Analysis*.

Throughout this assignment keep all decimals to four places, i.e. X.xxxx.

Any computations that involve "the log function", denoted by log(x), are always meant to mean the natural log function (which will show as ln() on a calculator). The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

When stating the null and alternate hypotheses in any statistical test in PREDICT 410, we should always state these hypotheses in terms of the model parameters, i.e. the model coefficients denoted by the betas.

Foundations of Logistic Regression:

(1) (5 points) What values can the response variable Y take in logistic regression, and hence what statistical distribution does Y follow?
The response variable Y is a binary variable that can take the values {0,1}, hence Y is assumed to have a binomial distribution.

(2) (5 points) How are the parameters estimated in logistic regression? Is this different from how the parameters are estimated in Ordinary Least Squares (OLS) regression?
The parameters in logistic regression are estimated by the method of maximum likelihood, or Maximum Likelihood Estimation (MLE). Typically we present the estimation of OLS models in terms of minimizing the sum of the squared error, but it turns out that minimizing the sum of the squared error for a regression model with Gaussian errors is equivalent to using maximum likelihood estimation. The primary difference in the estimation of the two models is that OLS regression has a Gaussian (Normal) likelihood function and logistic regression has a binomial likelihood function.

(3) (5 points) How do we define a "residual" in logistic regression, and how is it computed?
The residual of interest for logistic regression is the *deviance residual*. The OLS style residual e(i) = y(i) – yhat(i) is not a useful residual in logistic regression. The deviance residual is defined as r(i) = sign(y(i) – mu-hat(i))*sqrt(d(i)), where d(i) is the observation-specific component to the *model deviance*. The squared deviance residuals are the individual components of the model deviance. The difference between the deviance values for two nested models is a likelihood ratio test of the statistical significance of the additional predictor variables. Large deviance residuals represent observations that do not fit the model well, i.e. they are outliers.

Another valid choice would be the Pearson residuals, but I would like to steer you towards the deviance residual.

Model 1: Let's consider the logistic regression model, which we will refer to as Model 1, given by

$$\log(pi / [1-pi]) = 0.25 + 0.32*X1 + 0.70*X2 + 0.50*X3 \qquad (M1),$$

where X3 is an indicator variable with X3=0 if the observation is from Group A and X3=1 if the observation is from Group B. The likelihood value for this fitted model on 100 observations is 0.0850.

(4) (6 points) For X1=2 and X2=1 compute the log-odds for each group, i.e. X3=0 and X3=1.
For X3=0 log(pi / [1-pi]) = 0.25 + 0.64 + 0.70 = 1.59.
For X3=1 log(pi / [1-pi]) = 0.25 + 0.64 + 0.70 + 0.50 = 2.09.

(5) (6 points) For X1=2 and X2=1 compute the odds for each group, i.e. X3=0 and X3=1.
For X3=0 pi / [1-pi] = exp(1.59) = 4.9037.
For X3=1 pi / [1-pi] = exp(2.09) = 8.0849.

(6) (6 points) For X1=2 and X2=1 compute the probability of an event for each group, i.e. X3=0 and X3=1.
The probability of an event is computed by solving the odds ratio for pi. This solution yields the formula: pi = exp(XB) / [1 + exp(XB)].

For X3=0 pi = exp(1.59) / [1 + exp(1.59)] = 0.8306.
For X3=1 pi = exp(2.09) / [1 + exp(2.09)] = 0.8899.

(7) (2 points) Using the equation for M1, compute the relative odds associated with X3, i.e. the relative odds of Group B compared to Group A.
The relative odds of Group B to Group A is given by exp(0.5) = 1.6487.

(8) (5 points) Use the odds ratios for each group to compute the relative odds of Group B to Group A. How does this number compare to the result in Question #7. Does this make sense?
We can also compute the relative odds of Group B to Group A using the ratio of the odds ratios for each group, i.e. 8.0849 / 4.9037 = 1.6487. This is the same relative odds value, as it should be.

Model 2: Now let's consider an alternate logistic regression model, which we will refer to as Model 2, given by

$$\log(pi / [1-pi]) = 0.25 + 0.32*X1 + 0.70*X2 + 0.50*X3 + 0.1*X4 \quad (M2),$$

where X3 is an indicator variable with X3=0 if the observation is from Group A and X3=1 if the observation is from Group B. The likelihood value from fitting this model to the same 100 observations as M1 is 0.0910.

(9) (10 points) Use the G statistic to perform a likelihood ratio test of nested models for M1 and M2. State the hypothesis that is being tested, compute the test statistic, and test the statistical significance using a critical value for alpha=0.05 from Table A.3 on page 375 in *Regression Analysis By Example*. From these results should we prefer M1 or M2?

In this problem M2 nests M1 so M2 is the *full model (FM)* and M1 is the *reduced model (RM)*. The null hypothesis that we will be testing is that the parameter value for X4 is statistically equivalent to zero, i.e. H0: b4 = 0 versus H1: b4 != 0. For logistic regression the model deviance is given by D = -2*loglik and G = D(RM) – D(FM) (see pp. 13-14 in *Applied Logistic Regression*). Here we have D1 = -2*log(0.0850) = 4.9302 and D2 = -2*log(0.0910) = 4.7937 which implies that G = 4.9302 – 4.7937 = 0.1365. The statistic G follows a chi-squared distribution is the degrees-of-freedom equal to the difference in the number of model parameters between M1 and M2. In this case we have a chi-squared distribution with 1 degree-of-freedom. From Table A.3 for alpha = 0.05 we have a critical value of 3.84. Since 0.1365 < 3.84, we fail to reject the null hypothesis and conclude that we prefer M1 to M2.