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?

A logistic regression response variable is discrete taking on two or more variables. The response variable is modeled by the probability that the response will take a discrete value. The response variable follows 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 are estimated by the maximum likelihood function that maximize the probability of obtaining the observed set of data. In OLS regression the parameters are estimated by minimizing the sum-of-square deviations between the observed values and the predicted values. The least squares method yields estimators with a number of desirable statistical properties although when this same method is used for a discrete variable those same properties do not exist. Interestingly enough the maximum likelihood method that yields the least squares function is the same foundation for estimating parameters in logistic regression.

(3) (5 points) How do we define a "residual" in logistic regression, and how is it computed?

Deviance is the residual statistic in logistic regression similar to residual sum-of-squares for linear regression. For logistic regression, the residual is based on the ratio of the log-likelihood function of the fitted model and a saturated model (optimal). Since the saturated model contains as many parameters as data points, the ratio presents the fraction that is due to error. Specifically, deviance is calculated as:

D = 2ln(likelihood of the fitted model)

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 (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.

Group A: X3 = 0, Group B: X3=1

Log-odds Group A: g(x) = log ($\frac{pi}{1-pi}$) = 0.25 + 0.32*(2) + 0.70*(1) = 1.59

Log-odds Group B: g(x) = 0.25 + 0.32*(2) + 0.70*(1) + 0.50*(1) = 2.09

(5) (6 points) For X1=2 and X2=1 compute the odds for each group, i.e. X3=0 and X3=1.

Odds of outcome being present for Group A (X3=0) = $\frac{pi}{1-pi} = e^{g(x)} = e^{1.59}$ = 4.9037

Odds of outcome being present for Group B (X3-=1) = $e^{2.09}$ = 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.

Group A (X3=0): pi (x) = $\frac{e^{1.59}}{1+ e^{1.59}} = \frac{4.9037}{5.9037}$ = 0.8306 (also equivalent to: $\frac{O}{1+O} = \frac{4.9037}{1+ 4.9037}$ = $\frac{4.9037}{5.0937}$ =)

Group B (X3=1): pi (x) = $\frac{e^{2.09}}{1+ e^{2.09}} = \frac{8.0849}{9.0849}$ = 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.

$e^{0.50}$ = 1.6487

(8) (5 points) Use the odds 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?

Relative odds of Group B to Group A = $\frac{8.0849}{4.9037} = 1.6487$

These numbers are identical since when X_j is a binary variable, e^{B_j} is the actual odds ratio

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

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

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 example M2 is the full model (FM) and M1 is the reduced model (RM).

HO: X4 = 0
H1: X4 not equal to 0

G = - 2 ln $\left[\frac{(likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right]$ = -2 ln $(\frac{0.0850}{0.0910})$ = 0.1364

From Table A.3 the critical value is 124.34. Since 0.1364 is less than the critical value we cannot reject the null hypothesis. M1 is preferred.