

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

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

$$\log(\pi / [1-\pi]) = 0.25 + 0.32 \cdot X_1 + 0.70 \cdot X_2 + 0.50 \cdot X_3 \quad (M1),$$

where  $X_3$  is an indicator variable with  $X_3=0$  if the observation is from Group A and  $X_3=1$  if the observation is from Group B. The likelihood value for this fitted model on 100 observations is 0.0850.

- (4) (6 points) For  $X_1=2$  and  $X_2=1$  compute the log-odds for each group, i.e.  $X_3=0$  and  $X_3=1$ .
- (5) (6 points) For  $X_1=2$  and  $X_2=1$  compute the odds for each group, i.e.  $X_3=0$  and  $X_3=1$ .
- (6) (6 points) For  $X_1=2$  and  $X_2=1$  compute the probability of an event for each group, i.e.  $X_3=0$  and  $X_3=1$ .

- (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.
- (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?

**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?