EXST 7015
Fall 2014
Lab 9: Logistic Regression

OBJECTIVES

Logistic Regression is a type of predictive model that can be used when the dependent variable is a categorical variable with two categories – for example male/female, fail/pass, live/die, has disease/doesn't have disease, wins race/doesn't win, etc. Thus the dependent variable can take the value 1 with a probability of success (p), or the value 0 with a probability of failure (1-p). The independent or predictor variable can take any form (continuous, dichotomous and/or dummy variable with more than two categories). That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group.

The relationship between the independent and dependent variables is not a linear function as shown below:

$$\mathbf{p} = e^{(\alpha+\beta 1X1+ \beta 2X2 +\ldots+\beta iXi)}/\{1+ e^{(\alpha+\beta 1X1+ \beta 2X2 +\ldots+\beta iXi)}\}$$

where α = the constant of the equation and, β = the coefficient of the independent variables. The computed value, p, is a probability in the range of 0 to 1.

Much of the interpretation of logistic regression model centers on the ratio.

$$\mathbf{Odds} = p/(1-p) = e^{(\alpha+\beta 1X1+ \beta 2X2 +\ldots+\beta iXi)}$$

where Odds can take on values between zero and infinity.

The logarithm of odds, **logit,** results in a linear model, **the logistic regression:**

$$\mathbf{Logit} = \mathbf{log\ (odds)} = \log\{p/(1-p)\}= \alpha+\beta_1X_1+ \beta_2X_2 +\ldots+\beta_iX_i$$

Interpretation of the parameters differs in logistic regression, as parameter estimates need to be back-transformed to be meaningful. To estimate a predicted probability, you must calculate $\alpha+\beta_1X_1+ \beta_2X_2 +\ldots+\beta_iX_i$ at the desired X_is to get the predicted **logit** first, and then exponentiate it to get the predicted **Odds** that can be back-transformed as:

The predicted probability = predicted odds/(1+predicted odds)

The important tests generated by logistic regression are the "Tests of Global Null Hypothesis: Beta=0" and the "Analysis of Maximum Likelihood Estimates". The "Analysis of Maximum Likelihood Estimates" uses Wald statistics to test the null hypothesis H_0 that the

associated parameter estimates are not equal zero. The "Tests of Global Null Hypothesis" are essentially tests of model significance, much like the model F-test for linear regression. Typically, the best test to use is the likelihood ratio test, which uses a Chi-square test of significance to test whether the slope parameter βs are significant different from zero. If this test is not significant, it indicates that the logistic regression is not an appropriate model for the experimental data.

LABORATORY INSTRUCTIONS

Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78 ps=53;
title1 'EXST7015 lab 9, Name, Section#';
ods rtf  file = 'c:/temp/lab9.rtf';
ods html file = 'c:/temp/lab9.html';
```

Data set

The data set to be used is taken from a collection of data sets included in the textbook An Introduction to Categorical Data Analysis, written by Alan Agresti (John Wiley & Sons, Inc., New York, NY, 1996). The data consists only of data for the response (dependent) variable (a binary response of whether or not thermal distress was detected in a given O-ring seal on a space shuttle), and the explanatory (independent) variable (outside air temperature at time of shuttle launch), for a randomly and independently selected set of 23 shuttle launches. Two additional observations have been included in this dataset for purposes of estimating predicted responses to given values of the independent variable.

Fitting Logistic Model by Using PROC LOGISTIC

```
DATA one;
   input temp distress @@;
   cards;
66 0  70 1  69 0  68 0  67 0
72 0  73 0  70 0  57 1  63 1
70 1  78 0  67 0  53 1  67 0
75 0  70 0  81 0  76 0  79 0
75 1  76 0  58 1  66 .   35 .
;
proc print data=one;
run;
proc logistic data=one descending;
   Title2 'Logistic analysis of shuttle data';
   model distress = temp / CLPARM=wald CLODDS=wald alpha=0.01;
   output out=two p=predicted;
run;
proc print data=two;run;
proc sort;  by temp;  run;
proc gplot data=two;
Title3 'Plot of predicted prob vs. temp';
   plot predicted*temp;
   symbol c=blue i=join v=dot;
run;
```

Descending: Performing the logistic regression relative to "Success" probability, rather the "Failure" probability. Since it is desired to estimate the probability of thermal distress, the "Success" outcome, for different ambient temperatures, the **Descend** option is used in this lab. If this option is omitted, such that the "Failure" outcome is modeled, the probability of the "Success" outcome may still be estimated by solving the odds ratio algebraically.

CLPARM=wald and **CLODDS**=wald: Calculating confidence limits for the estimates of the model parameters and associated odds ratio. Confidence limits are difficult to calculate manually, so it is easy to specify the option to do so automatically. Both sets of confidence limits are calculated using a **Wald test statistics**, which essentially is a Z-test (calculated by parameter estimate divided by the standard error, $Z=\beta/SE$)

LAB ASSIGNMENT

1. Write the logistic regression equation to model the odds of distress as a function of temperature.

2. Perform a logistic regression, and report the regression parameters and their 99% confidence intervals.

3. Does temperature affect the odds of distress? Explain the reason for your answer.

4. What is the probability of distress at 66 degrees and at 35 degrees?

5. Plot the probability curve and describe it.