Name: _____

1 Objective: Comparing Proportions and Logistic Regression

What differentiates a logistic regression model from others is that the outcome variable in logistic regression is binary (or dichotomous or two levels, e.g.,yes/no, sick/well, pass/fail etc.). When you have a dependent variable with only two levels multiple-regression techniques are not appropriate. Logistic regression uses a transformation (called a `logit`) which forces the prediction equation to predict values between 0 and 1. A logistic regression equation predicts the natural log (ln) of the odds for a subject being in one category or another. In addition, the regression coefficients in a logistic regression equation can be used to estimate odds ratios for each of the independent variables (read C.S. 300-315). Logistic function is given by

$$y = \frac{L}{1 + e^{a+bx}}$$

where a, b, and L (carrying capacity or horizontal asymptote) are constants. For different values of a and b, logistic function generates a variety of S-shaped curves. The form that is useful for us is is called *logit transformation* and give by

$$\ln\left(\frac{L-y}{y}\right) = a + bx$$

which implies that $\ln\left(\frac{L-y}{y}\right)$ is linear with x. Here $\frac{L-y}{y}$ refers to odds and logarithm of odds refers to `logit` or log-odds.

In general, the logistic model for a binary dependent variable can be written as

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}$$

where

$$y = \begin{cases} 1 & \text{if category A occurs} \\ 0 & \text{if category B occurs} \end{cases}$$

$$E(y) = P(\text{if category A occurs}) = \pi$$

$x_1, x_2, \cdots, x_n$ are quantitative or qualitative independent variables. As in single variable case,

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Let $\pi^* = \ln\left(\frac{\pi}{1-\pi}\right)$. The transformed logistic model

$$\pi^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

is linear in the β's and the method of least squares can be applied. Note that, since $\pi = P(y = 1)$, then $1 - \pi = P(y = 0)$. The ratio

$$\frac{\pi}{1-\pi} = \frac{P(y=1)}{P(y=0)}$$

is known as the *odds* of the event, $y = 1$, occurring.

2 Key Words

`proc logistic, oddsratio, effectplot, tables`, categorical data, logistic modeling.

3 Lab Steps

We will consider the heart attack data stored in `risk.xlsx` for this lab (SAS Statistics by R.Cody).

[A] We will **compare proportions** using chi-square test. We can use this test with data in which each observation represents a single subject or with data that consists of frequency counts. Descriptions that we are interested in are relative risks and odds ratios.

[1] Import data `risk.xlxs` into SAS (proc import) and identify variables "Age_Group (1=<60, 2=60-70, 3=71+), Gender (M=Male, F=Female), Age (in years), Chol (Cholesterol level), Heart_Attack (1=Yes, 0=No).

[2] Run the frequency procedure for the genders: Outputs 2x2 table with Gender as the rows and Heart_Attack as the columns.

```
TITLE "Comparing Proportions";
PROC freq data=risk;
   TABLES Gender * Heart_Attack / chisq;
run;
```

[3] General form of a two-way table has the form:
`tables row-variable * column-variable / options;`
Options include CHISQ, FISHER, MEASURES, and few more.

[4] Note the followings for females:

- There were 233 females who did not have a heart attack.
- Of the 500 people in this study, 46.60% of them were females who did not have a heart attack.
- Of all the females, 93.20% did not have a heart attack.
- Of the 442 people who did not have a heart attack, 52.71% of them were female.

[5] List the above proportions for males who had a heart attack:

- Number of males who had a heart attack _____
- Percentage of males who had a heart attack (out of the 500 people) _____
- Of all the males, percentage of males who had a heart attack _____
- Among the people who had a heart attack, percentage of male _____

[6] Determine whether this difference in proportions is statistically significant: _____

_____

_____

_____

[7] Rearranging Rows and Columns in a Table and Computing Relative Risk: Run the code below and compare the output with the output of [2].

```
PROC format;
   value $gen 'M' = '1:Male'
              'F' = '2:Female';
   value attack 1 = '1:Yes'
                0 = '2:No';
run;

TITLE "Reordering the Rows and Columns in a 2x2 Table";
PROC freq data=risk order=formatted;
   tables Gender * Heart_Attack / chisq relrisk;
   format Gender $gen. Heart_Attack attack.;
run;
```

[8] Output shows the relative risk and odds ratio: For cohort study, the relative risk for a man having a heart attack compared to that of a woman would be 2.4118 (95%CI 1.4098, 4.1284). Because the 95% CI does not include 1 (ratio of M/F heath attack, i.e. number of the heath attack of M and F are the same), we would report this as a statistically significant result at the .05 level. The last row of output is labeled Col2 Risk. This is the relative risk for men not having a heart attack (column 2).

	1-Yes	**2-No**
M	41	209
F	17	233

Cohort (Col1 Risk)	41/17
Cohort (Col2 Risk)	209/233

For a case-control study instead of a cohort study, you would report the odds ratio as 2.6887, Row1/Row2 (with a 95% CI of 1.4824 to 4.8768).

[B] Logistic regression with one categorical predictor variable:

[1] Read `risk.dat` into SAS and identify variables "Age_Group (1=<60, 2=60-70, 3=71+), Gender (M=Male, F=Female), Age (in years), Chol (Cholesterol level), Heart_Attack (1=Yes, 0=No).

[2] Run the logistic regression below (note that predictor varibale is categorical (gender)):

```
TITLE "Logistic Regression with One Categorical Predictor Variable";
PROC logistic data=risk;
   class Gender (param=ref ref='F');  /* causing females to be the reference level */
   model Heart_Attack (event='1') = Gender / clodds = pl;
run;
quit;
```

[3] `CLODDS=PL` option indicates that we want SAS to compute confidence limits on the odds ratio using the method of profile likelihood (PL). PL is used more than 50 data sample size. Any sample size less than 50, it's better to use Wald intervals (Allison, Paul D. 1999, Logistic Regression Using SAS).

[4] Note that the log odds of having a heart attack is equal to -2.6178 + .9890 x Gender ($\ln y = -2.6178 + .9890 \times \text{Gender}$), where Gender is 0 for females and 1 for males. Log-odds of heart attack for females is -2.6178 ($e^{-2.6178} = 0.07296$). The difference in log-odds is expected to be 0.9890 ($e^{0.9890} = 2.689$) units higher for males compared to females.

Because females are the reference level, this odds ratio means that the odds of a male having a heart attack are 2.689 times that for a female. Find these values in the output.

[5] In this part, we will use cholesterol level (continuous variable Chol) as a predictor of heart attack:

[6] Run the logistic regression below:

```
TITLE "Logistic Regression with One Continuous Predictor Variable";
PROC logistic data=risk;
        model Heart_Attack (event='1') = Chol / clodds = pl;
        units Chol = 10; /* odds ratios are calculated per 10 unit increase of Chol */
run;
quit;
```

Note that there's no CLASS statement, since this model does not have categorical variables. Without a UNITS statement, the odds ratios shown in the output would represent the increase in the odds of having a heart attack for each unit increase in cholesterol. In this program, you are asking for the odds ratio for each increase of 10 units. You can include as many variables and units as you want on one UNITS statement.

[7] **Examine the output to see if cholesterol level is significant in determining heart attack:** Log-odds of having a heart attack is equal to -2.6178 + .9890 x Gender ($\ln y = -5.9979 + 0.0192 \times \text{Chol}$), i.e., 1 unit increase in cholesterol increases odds ratio of having heart attack by $e^{0.0192} = 1.019$ unit; 10 unit increase in cholesterol increases odds ratio of having heart attack by $e^{0.192} = 1.212$ unit.

[8] Add `plots(only) = effect;` after `PROC logistic data=risk` in the previous code. What do you observe? Explain.

[9] Categorizing the continuous variable `Chol`: We want to create two categories for cholesterol: 'Low to Medium' and 'High', using CLASS variables.

```
PROC format;
   value cholgrp low-200   = 'Low to Medium'
                 201-high  = 'High';
run;

TITLE "Using a Format to Create a Categorical Variable";
PROC logistic data=risk;
   class Chol (param=ref ref='Low to Medium');
   model Heart_Attack (event='1') = Chol / clodds = pl;
   format Chol cholgrp.;
run;
quit;
```

[10] Examine the output in this model to see if cholesterol level is significant in determining heart attack.

[11] Using a Combination of Categorical and Continuous Variables in a Logistic Regression Model (note the predictor multivariables):

```
ods graphics on; /* to obtain a default plot that shows the odds ratios for each predictor */
TITLE "Using a Combination of Categorical and Continuous Variables";
PROC logistic data=risk;
        class Age_Group (ref='1:<_60')
                Gender (ref='F') / param=ref;
        model Heart_Attack (event='1') = Gender Age_Group Chol / clodds = pl;
        units Chol=10;
run;
quit;
ods graphics off;
```

[12] Interpret the profile likelihood odds ratios and the graph displayed in the output.

[C] Logistic regression with interactions: In `PROC logistic`, we will specify main effects and possible two-way interactions in the model.

[1] Note the interactions of each predictor in the model below and Receiver Operating Characteristic (ROC) plot request.

```
ods graphics on;
TITLE "Running a Logistic Model with Interactions";
PROC logistic data=risk plots(only)=(roc oddsratio);
        class Gender (param=ref ref='F')
              Age_Group (param=ref ref='1:<60');
        model Heart_Attack (event='1')=Gender|Age_Group|Chol @2/selection=backward slstay=.10
            clodds=pl;
        units Chol=10;
        oddsratio Chol;
        oddsratio Gender;
        oddsratio Age_Group;
run;
quit;
ods graphics off;
```

Note that `@2`, after the model specification, limits the model to have no higher than two-way interactions.

[2] **ROC curve:** It shows the relationship between a false-positive rate and the sensitivity of the test. Ideally, models have high sensitivity and low false-positive rates. An ROC curve gives you a visual representation of the relationship between the falsepositive rate and the sensitivity.

[D] Turn In: How much the model is capable of distinguishing between classes? What is the cut off `chol` point in determining heart attack? Which predictor is significant in determining heart attack? Explain your results and graphs.