STA 312f10 Assignment 10

Do this assignment in preparation for the quiz on Friday, Nov. 26th. Please bring your SAS log file and list file to the quiz; they may be handed in. Please do *not* write anything on your printouts before the quiz, except possibly your name and student number.

1. This question deals with the the simplest case of the Wald test, where the parameter is one-dimensional. You may need the some (not all) of the following formulas, which also will appear on the Final Exam formula sheet.

   $$\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})_{k\times k} = \left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\boldsymbol{\theta},\mathbf{Y})\right] \qquad \widehat{\mathbf{V}} = \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}})^{-1}$$

   $$W = (\mathbf{C}\widehat{\boldsymbol{\theta}}-\mathbf{h})'(\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}}-\mathbf{h}) \quad S = \mathbf{u}(\widehat{\boldsymbol{\theta}}_0)'\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}_0)^{-1}\mathbf{u}(\widehat{\boldsymbol{\theta}}_0),\ \mathbf{u}(\boldsymbol{\theta})_{k\times 1} = \left[\frac{\partial\ell}{\partial\theta_i}\right]$$

   Remember, for one-dimensional problems like this, matrix multiplication is scalar multiplication, transposing a scalar does nothing, and $x^{-1} = \frac{1}{x}$.

   Let $X_1, \ldots, X_N$ be a random sample from a $B(1,\theta)$ distribution. We want to test $H_0 : \theta = \theta_0$ several ways. You may take it as given that the log likelihood is $\ell(\theta) = x\log\theta + (N-x)\log(1-\theta)$, where $x = \sum_{i=1}^N x_i$, and that the unrestricted MLE is $\widehat{\theta} = x/N$. You could easily do these calculations.

   (a) Find $\mathcal{J}(\theta)$, the Fisher information. It is a function of the parameter θ.

   (b) Substitute $\theta = \widehat{\theta}$ to obtain $\mathcal{J}(\widehat{\theta})$. Simplify!

   (c) Find $\widehat{V}$. It is a function of x and N.

   (d) Show that the Wald test statistic is

   $$W = \frac{N(\widehat{\theta}-\theta_0)^2}{\widehat{\theta}(1-\widehat{\theta})}.$$

   (e) Now suppose a sample of 100 consumers is asked to choose between 2 kinds of soap, and say which one works better. Sixty choose Brand A and 40 choose Brand B. What is the critical value at $\alpha = 0.05$ for *all* your Chisquare tests? The answer is a number. A table will be supplied on the quiz an Final Exam if necessary.

   (f) What is the value of the Wald test statistic for the soap data? Your answer is a number.

   (g) Do you reject H_0 with the Wald test? Answer Yes or No.

   (h) What is the value of the Likelihood Ratio test statistic for the soap data? Your answer is a number. Hint: What are the expected frequencies under H_0?

   (i) Do you reject H_0 with the Likelihood Ratio test? Answer Yes or No.

   (j) What is the value of the Pearson Chisquare test statistic for the soap data? Your answer is a number.

   (k) Do you reject H_0 with the Pearson Chisquare test? Answer Yes or No.

   (l) In plain, non-statistical language, what do you conclude from these tests? Your answer is about which kind of soap is perceived to be better.

(m) Suppose you were to set this up as a logistic regression with no independent variables. Thus, your only parameter is β_0. In terms of β_0, what is the null hypothesis?

(n) In terms of β_0, what is the likelihood function? It is a function of β_0, x and N. Simplify as much as possible.

(o) What is the maximum likelihood estimator $\widehat{\beta}_0$? Your answer is a symbolic expression in x and N. Show your work.

(p) What is the maximum likelihood estimate $\widehat{\beta}_0$ for the soap data? Your answer is a number.

(q) As a symbolic expression in x and N, what is $\ell(\widehat{\beta}_0)$? Compare this to $\ell(\widehat{\theta})$, also written as a function of x and N.

(r) Do you realize that your symbolic $\ell(\widehat{\beta}_0)$ applies to *any* logistic regression model with no independent variables? Answer Yes or No. Thus, if you have $-2\log\ell(\widehat{\boldsymbol{\beta}})$ for any logistic regression model, all you need is the number of $y = 1$ cases and the sample size in order to calculate a likelihood ratio test of all the variables at once. SAS calls this the "`Global Null Hypothesis:  BETA=0`."

2. The Data Sets link on the course web site will lead you to the Heart Data. It comes from a sample of middle-aged men who worked for the Western Electric Company in the late 1950s. Use `curl` to get a copy of `heart.txt`, and also `heartread.sas`.

You will save a lot of time and trouble by using `%include 'heartread.sas'` in your SAS program. Note that `heartread.sas` reads the data directly from `heart.txt`, skipping all the lines at the top that describe the data set.

For this assignment, the response variable is heart attack. Please confine your analyses to the following variables: `age diastol cholest bmi smoker famhist educat attack`. Body Mass Index (`bmi`) is a measure of how heavy someone is relative to his or her height. Values over 25 are supposed to mean that the person is overweight.

(a) First, obtain means and standard deviations of the continuous variables, and frequency distributions of the categorical variables.

(b) Now please consider a very simple logistic regression model with just one independent variable: education (the categorical version, of course).

i. Are the independent and dependent variables related? You have at least 3 tests (some appearing more than once) that seek to answer this question. For each one, be able to give the value of the test statistic, the degrees of freedom, the p-value (all numbers), and whether or not the null hypothesis is rejected at the 0.05 significance level.

ii. Is there a meaningful difference in the chances of heart attack between men with a grade school education and university graduates? Give the Wald chisquare statistic, the degrees of freedom, the p-value, and answer the question Yes or No; be guided by the 0.05 significance level.

iii. The estimated odds of a heart attack are ____ times as great for a man with a grade school education, compared to a university graduate.

iv. Is there a meaningful difference in the chances of heart attack between men with a high school education and university graduates? Give the Wald chisquare statistic, the degrees of freedom, the p-value, and answer the question Yes or No; be guided by the 0.05 significance level.

v. The estimated odds of a heart attack are ___ times as great for a man with a high school education, compared to a university graduate.

vi. Is there a meaningful difference in the chances of heart attack between men with a some college and university graduates? Give the Wald chisquare statistic, the degrees of freedom, the p-value, and answer the question Yes or No; be guided by the 0.05 significance level.

vii. The estimated odds of a heart attack are ___ times as great for a man with some college, compared to a university graduate.

viii. There are three more comparisons like this; please carry them out. If you used the same dummy variable coding scheme I did, you will have to request special Wald tests. If you are still following this, you may be wondering about the names of dummy variables when there are embedded blanks. Try replacing them with underscores.

ix. In plain language, what do you conclude from all these tests? Try to make it short and sweet.

x. Using numbers from your printout and a calculator, estimate the probability of a heart attack for study participants with a high school education.

xi. Use `proc freq` to check your likelihood ratio test, some of your odds ratios (try $\widehat{\alpha}$), and the requested probability estimate.

(c) Now add Body Mass Index to the model. Please use Wald tests to answer the following questions:

i. Controlling for `bmi`, is education significantly related to heart attack? Answer Yes or No and give the value of the test statistic and the p-value (numbers from the printout).

ii. Controlling for education, is `bmi` significantly related to heart attack? Answer Yes or No and give the value of the test statistic and the p-value (numbers from the printout). State your conclusion, if any, in plain language.

iii. Give the value of the test statistic and the p-value for the simultaneous test of `bmi` and education. Again, these are numbers from the printout.

iv. Controlling for education, for each unit of increase in body mass index, the estimated odds of coronary heart disease are multiplied by ___. The answer is a number from your printout. Please disregard the significance test this time.

(d) Using all available variables in the list provided, find a good model for predicting heart attack. I did it informally (no automatic methods) and came up with an interesting two-variable model pretty fast. We'll pursue this in the final assignment and the final exam.

3. The SAS logistic regression list files contain some redundant information. If you know what is going on, you can figure out what some of the numbers have to be, based on other information on the printout. At the end of this document is a short list file with some of the numbers underlined. On a quiz or the final they would be blanks. Your job is to calculate them based on other numbers in the list file. The green ones are pretty quick. The two red ones are more challenging. There is a little rounding error in some cases.

Fill in the blanks

```
                            The LOGISTIC Procedure

                               Model Information

      Data Set                      WORK.HEART
      Response Variable             attack                Had Heart Attack
      Number of Response Levels     2
      Model                         binary logit
      Optimization Technique        Fisher's scoring


                    Number of Observations Read         239
                    Number of Observations Used         238


                                Response Profile

                        Ordered                      Total
                          Value     attack       Frequency

                              1     No                 121
                              2     Yes                117

                     Probability modeled is attack='Yes'.

NOTE: 1 observation was deleted due to missing values for the response or
      explanatory variables.


                           Class Level Information

                                               Design
                        Class      Value     Variables

                        smoker     0                 1
                                   1                -1


                           Model Convergence Status

                Convergence criterion (GCONV=1E-8) satisfied.


                             Model Fit Statistics

                                                 Intercept
                                  Intercept            and
                    Criterion          Only     Covariates

                    AIC             331.871        324.868
                    SC              335.343        335.285
                    -2 Log L        329.871        318.868
```

```
                   Western Electric (Heart) Study: STA312f10                 10
                             SA312f10 HW 10 Check
                             Fill in the blanks!

                            The LOGISTIC Procedure

                    Testing Global Null Hypothesis: BETA=0

            Test                 Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio        11.0031        2         0.0041
            Score                   10.8035        2         0.0045
            Wald                    10.4217        2         0.0055

                          Type 3 Analysis of Effects

                                          Wald
                  Effect      DF    Chi-Square    Pr > ChiSq

                  age          1        4.4616        0.0347
                  smoker       1        6.8927        0.0087

                   Analysis of Maximum Likelihood Estimates

                                      Standard          Wald
     Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq

     Intercept       1     -3.3895      1.5710        4.6552        0.0310
     age             1      0.0690      0.0327        4.4616        0.0347
     smoker    0     1     -0.3548      0.1352        6.8927        0.0087

                             Odds Ratio Estimates

                                   Point          95% Wald
               Effect           Estimate      Confidence Limits

               age                 1.071       1.005       1.142
               smoker 0 vs 1       0.492       0.290       0.835

         Association of Predicted Probabilities and Observed Responses

               Percent Concordant     59.8    Somers' D    0.230
               Percent Discordant     36.8    Gamma        0.238
               Percent Tied            3.4    Tau-a        0.115
               Pairs                 14157    c            0.615

                      Linear Hypotheses Testing Results

                                     Wald
                Label          Chi-Square       DF    Pr > ChiSq

                MysteryTest       10.4217        2        0.0055
```