Homework Assignment 3

Due Date: Monday, Feb 25, 2019 at 1PM

Total Points: 75

Please hand in a print-out of your answer and R code, and also email your R code to me (hoyen@stat.sc.edu). Use R markdown for Questions 2. You can hand-write Question 1, 3.

1 Hypothesis Testing

(35 points) There are three frequently occurring test statistics, the likelihood ratio test, the Wald test, and the score test. If $\mathbf{Y}$ has the probability density function $f(y|\boldsymbol{\beta})$ at $\mathbf{Y} = y$, where $\boldsymbol{\beta}$ is $p \times 1$, then hypothesis of interest are often of the form $H_0 : \mathbf{L}'\boldsymbol{\beta} = \xi$ versus $H_1 : \mathbf{L}'\boldsymbol{\beta} \neq \xi$, where $\mathbf{L}'$ is $s \times p$ of rank $s < p$. Let

- $\widehat{\boldsymbol{\beta}}$ denotes the MLE of $\boldsymbol{\beta}$ under the full model.
- $\tilde{\boldsymbol{\beta}}$ denotes the MLE of $\boldsymbol{\beta}$ under the model assuming the null hypothesis is true,
- $\ell(\boldsymbol{\beta}) = \log\left[f(y|\boldsymbol{\beta})\right]$ denote the log likelihood function,
- $s(\boldsymbol{\beta})$ be the vector of score with j^{th} component, $s_j(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}$
- $I(\boldsymbol{\beta})$ be Fisher's information matrix which has j, k element equal to $-E[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial^2 \boldsymbol{\beta}_j \boldsymbol{\beta}_k}]$.

The three test statistics in this case are:

- Likelihood ration test statistics: $-2[\ell(\tilde{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}})]$
- Wald test statistic: $(\mathbf{L}'\widehat{\boldsymbol{\beta}} - \xi)'[\mathbf{L}'I(\widehat{\boldsymbol{\beta}})^{-1}\mathbf{L}]^{-1}(\mathbf{L}'\widehat{\boldsymbol{\beta}} - \xi)$
- Score test statistic: $s'(\tilde{\boldsymbol{\beta}})I(\tilde{\boldsymbol{\beta}})^{-1}s(\tilde{\boldsymbol{\beta}})$

For the logistic regression model $Y \sim Bernoulli(\frac{e^{\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2}}{1+e^{\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2}})$, where $\mathbf{X}_1$ is $n \times q$ of rank q, $\mathbf{X}_2$ is $n \times (p-q)$, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ is $n \times p$ of rank p. Derive the three test statistics for test $H_0 : \boldsymbol{\beta}_2 = 0$ versus $H_1 : \boldsymbol{\beta}_2 \neq 0$. Test statistics should be expressed in matrix form (e.g. written as a product of the matrices/vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{Y}$) and reduced as much as possible. Comment.

2 IRLS Algorithm

(25 points) Based on the logistic regression model described Question 1,

(a) Simulated data

(b) Write your own IRLS algorithm to produce the estimates of regression coefficients, standard error, test statistics for regression coefficients, and p-values.

(c) Compare your result with output from **R**. They should be the same.

3 Connection of logistic regression to 2 $\times$ 2 tables

Use the Medical Expenditure Panel Survey (MEPS) dataset for the following analysis. The MEPS data is available at `http://people.stat.sc.edu/hoyen/Stat704-2018/Data/h129.RData` and the codebook at `https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_codebook.jsp?PUFId=H129&sortBy=Start`

(a) Make a 2 $\times$ 2 table of mscd and smoking status. Calculate the log odds ration, its standard error and 95% CI using methods for 2 $\times$ 2 tables. To simplify the analysis, drop those people who have missing value of mscd and smoking status (this is to simplify the exercise but in practice is not generally a good strategy.

(b) Logistic regress mscd (Y) on smoking status (X). Compare the regression coefficient and its standard error with he log odds ratio and standard error calculated in 3(a).

(c) Logistic regress smoking status (Y) on mscd (X). Compare the regression coefficient and its standard error with he log odds ratio and standard error calculated in 3(a) and 3(b).

(d) Review the paper by Prentice and Pyke (Biomtrika , 1979) and then state the invariance property of the log odds ratio estimate from a logistic regression in precise mathematical terms.

4 Reference

1. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979;66:403.