Homework Assignment 3

Due Date: Friday, March 03, 2023 at 5PM

Total Points: 90

Please email your answer (compiled pdf file from R markdown) and R code to Anderson Bussing (`ABUSSING@email.sc.edu`). Please use the R markdown Homework template (Stat705_HWtemplate.Rmd) to write your homework solutions. You can hand-write Question 1 (a), 2, 4.

1 ROC Curves

(30 points) You are asked to evaluate the performance of two classification models, M_1 and M_2. The test set you have chosen contains 10 covariates, labeled as $A_1, ..., A_{10}$. Table 1 below shows the posterior probability obtained by applying the models to the test set. Assume that we are mostly interested in detecting the positive samples.

Sample	True Class	$P(+\|A_1, A_2, ..., A_{10}, M_1)$	$P(+\|A_1, A_2, ..., A_{10}, M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

Table 1: Table 1: Posterior probabilities by applying M_1, and M_2 on the test set

(a) (10 points) Write R code to draw ROC curve for both M_1 and M_2 in the same graph. Which model do you think is better? Explain your reasons. Manually show the steps for drawing ROC curve for M_1.

(b) (10 points) For model M_1, suppose you choose the cutoff threshold to be $P(+|A_1, A_2, ..., A_{10}, M_1) = 0.5$. In other words, any sample whose posterior probability is greater than 0.5 will be classified as a positive case. Compute the sensitivity and specificity.

(c) (10 points) Repeat part (b) for Model M_1 using threshold $P(+|A_1, A_2, ..., A_{10}, M_1) = 0.1$. Which threshold do you prefer 0.5 or 0.5? Are the results consistent with what you expect from the ROC curves?

2 Hypothesis Testing (Bonus Question)

(Bonus: 30 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.

3 IRLS Algorithm

(30 points) Write down a simple logistic regression model and

(a) Simulate 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.

4 Connection of logistic regression to 2 × 2 tables

(30 points) 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/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 2×2 table of mscd and smoking status. Calculate the log odds ration, its standard error and 95% CI using methods for 2×2 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.

5 Reference

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