Statistical Methods - Homework #2

1. Consider the logistic regression model

$$\log \frac{p_i}{1-p_i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad y_i \sim \text{Bernoulli}(p_i) \ \text{ independently}, \quad i = 1, \cdots, n.$$

(A) Show that

$$p_i = \frac{1}{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})}$$

Also, find $\dfrac{\partial p_i}{\partial \boldsymbol{\beta}}$.

(B) The likelihood $L(\boldsymbol{\beta})$ and the log-likelihood $\ell(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ are

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n p_i^{y_i} (1-p_i)^{1-y_i}, \quad \ell(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \log p_i + (1-y_i) \log(1-p_i)$$

respectively. Find $S(\boldsymbol{\beta}) = \dfrac{\partial \ell}{\partial \boldsymbol{\beta}}$ and $\dfrac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}$.

(C) The maximum likelihood estimator(MLE) $\hat{\boldsymbol{\beta}}$ solves $S(\hat{\boldsymbol{\beta}}) = 0$. Find the asymptotic distribution of the MLE $\hat{\boldsymbol{\beta}}$.

(D) Consider the hypothesis test:

$$H_0 : \boldsymbol{\beta} = 0, \quad H_A : \boldsymbol{\beta} \neq 0.$$

Show that the Wald test statistic is

$$W = \sum_{i=1}^{n} \hat{p}_i(1 - \hat{p}_i) \left(\log \frac{\hat{p}_i}{1 - \hat{p}_i} \right)^2,$$

and the likelihood ratio statistic is

$$\Lambda = 2 \sum_{i=1}^{n} \left\{ y_i \log(2\hat{p}_i) + (1 - y_i) \log(2(1 - \hat{p}_i)) \right\},$$

where $\hat{p}_i = p_i(\hat{\boldsymbol{\beta}})$. What are the null distribution of these test statistics?

2. Consider the logistic regression model with a single binary covariate $x_i \in \{0, 1\}$:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i, \quad y_i \sim \text{Bernoulli}(p_i) \text{ independently}, \quad i = 1, \cdots, n.$$

The data are summarized in the following contingency table:

		Y		
		0	**1**	**Total**
X	0	n_{00}	n_{01}	$n_{00} + n_{01} = n_0$
	1	n_{10}	n_{11}	$n_{10} + n_{11} = n_1$
	Total	$n_{00} + n_{10}$	$n_{01} + n_{11}$	$n_{00} + n_{01} + n_{10} + n_{11}$

(A) For subjects with $x_i = 0$, $p_i = p(0)$; for $x_i = 1$, $p_i = p(1)$, where $p(0) = \dfrac{e^{\beta_0}}{1 + e^{\beta_0}}$ and $p(1) = \dfrac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$.

Express the odds ratio(OR)

$$\text{OR} = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}$$

in terms of β_0 and β_1.

(B) Denote $n_0 = n_{00} + n_{01}$ and $n_1 = n_{10} + n_{11}$. Using the score equations $S(\boldsymbol{\beta}) = \mathbf{0}$ from 1.(B), show that the MLEs of $p(0)$ and $p(1)$ are

$$\hat{p}(0) = \frac{n_{01}}{n_0}, \qquad \hat{p}(1) = \frac{n_{11}}{n_1}.$$

(C) Using the result of (B), find the closed form MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$.

(D) Let $q(j) = 1 - p(j)$ for $j = 0, 1$. Using the result of 1.(B) with $\boldsymbol{x}_i = (1, x_i)^\top$, show that the Fisher information matrix is

$$\mathcal{I}(\boldsymbol{\beta}) = \begin{pmatrix} n_0 p(0) q(0) + n_1 p(1) q(1) & n_1 p(1) q(1) \\ n_1 p(1) q(1) & n_1 p(1) q(1) \end{pmatrix}.$$

Then find the asymptotic variance of $\hat{\beta}_1$, and evaluate it at the MLE.

(E) Find a test statistic and its null distribution for the following hypothesis test:

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0.$$

3. (Linear Discriminant Analysis) Show that finding $k \in \{1, \cdots, K\}$ that maximizes

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu_k)^2\}}{\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu_\ell)^2\}}$$

is equivalent to finding $k \in \{1, \cdots, K\}$ that maximizes

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

Also, when $K = 2$, show that this is equivalent to choosing $k = 1$ if

$$\frac{(\mu_2 - \mu_1)}{\sigma^2}\left(x - \frac{\mu_1 + \mu_2}{2}\right) < \log\left(\frac{\pi_1}{\pi_2}\right).$$

4. Consider the following confusion matrix and answer the following questions.

		True	
		Negative	**Positive**
Predicted	**Negative**	75	20
	Positive	25	80

(1) Compute the False Positive Rate (FPR).

(2) Compute the False Negative Rate (FNR).

(3) Compute the Sensitivity (True Positive Rate).

(4) Compute the Specificity (True Negative Rate).