Homework 3

Jieqi Tu

3/30/2021

5.3

(a) Construct the ROC curve for the toy example in Section 5.4.2. with complete separation.

```
# data import
x = c(1, 2, 3, 4, 5, 6)
y = c(1, 1, 1, 0, 0, 0)

# fit model
toy.model = glm(y~x, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
pred = prediction(fitted(toy.model), y)
```

5.14

Assuming $\pi_1 = \pi_2 = \cdots = \pi_N = \pi$, then the log likelihood would be

$$L(\pi) = \sum_{i=1}^{N} y_i log(\pi) + (n_i - y_i) log(1-\pi)$$

Take the first derivative, we can get

$$L'(\pi) = \frac{\sum y_i}{\pi} - \frac{\sum n_i - y_i}{1-\pi}$$

Set it equal to 0, we can get

$$\hat{\pi} = (\sum y_i)/(\sum n_i)$$

And the second derivative of $L(\pi)$ also confirms that $\hat{\pi}$ maximizes the likelihood function. Then the Pearson statistic for ungrouped data (when $ n_i=1$) is:

$$\begin{aligned}
\chi^2 &= \sum \frac{(observed - fitted)^2}{fitted} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{(y_{ij} - \hat{\pi})^2}{\hat{\pi}} + \frac{[1 - y_{ij} - (1-\hat{\pi})]^2}{1-\hat{\pi}} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{(y_{ij} - \hat{\pi})^2}{\hat{\pi}(1-\hat{\pi})} \\
&= \frac{N\hat{\pi}(1-\hat{\pi})}{\hat{\pi}(1-\hat{\pi})} = N
\end{aligned}$$

Since the Pearson statistic $\chi^2 = N$, the statistic is not informative for us to test the goodness-of-fit of the null model.

5.15

The log likelihood is $\sum_i [y_i log\pi_i + (1-y_i)log(1-\pi_i)]$. For the saturated model, we have $\hat{\pi}_i = y_i$ and the value of the log likehood of saturated model equals 0 (because y_i can only take value of 0 and 1).

$$\begin{aligned}
D(y;\hat{\boldsymbol{\mu}}) &= -2\sum observed \times log(observed/fitted) \\
&= -2(\sum_i y_{ij} log(\frac{y_i}{\hat{\pi}_i}) + \sum_i (1-y_i)log(\frac{1-y_i}{1-\hat{\pi}})) \\
&= -2\sum_i [y_i log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) + log(1-\hat{\pi}_i)] \\
&= -2\sum_i [y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) + log(1-\hat{\pi}_i)]
\end{aligned}$$

From 5.14, we could know that $\sum_i y_i = \sum_i \hat{\pi}_i$ and so $\sum_i x_i = \sum_i x_i\hat{\pi}_i$. So the deviance would be:

$$\begin{aligned}
D &= -2[\hat{\beta}_0 \sum_i \hat{\pi}_i + \hat{\beta}_1 \sum_i x_i\hat{\pi}_i + \sum_i log(1-\hat{\pi}_i)] \\
&= -2[\sum_i \hat{\pi}_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) + \sum_i log(1-\hat{\pi}_i)] \\
&= -2\sum_i \hat{\pi}_i log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) - 2\sum_i log(1-\hat{\pi}_i)
\end{aligned}$$

Therefore, the deviance only depends on $\hat{\pi}_i$, and it is uninformative for checking model fit.