(1) Suppose for the ith subject we observe x_i and y_i. Let $p(x_i;\beta) = P(Y=1|X=x_i)$.

Maximum likelihood estimation:

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{N} \Big\{ y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta)) \Big\} \\ &= \sum_i \Big\{ y_i x_i^\top \beta - \log\left(1+\exp(x_i^\top \beta)\right) \Big\}\end{aligned}$$

Please derive the blue part.

[proof] $l(\beta) = \sum_{i=1}^{N} \{ y_i \log p(x_i;\beta) + (1-y_i) \log(1-p(x_i;\beta)) \}$

$= \sum_{i=1}^{N} \{ y_i X^\top\beta - y_i \log(1+e^{X^\top\beta}) - (1-y_i)\log(1+e^{X^\top\beta}) \}$

$= \sum_{i=1}^{N} \{ y_i x^\top\beta - \log(1+\exp(x^\top\beta)) \}$

(2) Write Newton-Raphson algorithm to estimate logistic regression.

Reminder: you need to derive the equation

$$\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^\top} = -\sum_i x_i x_i^\top p(x_i;\beta)\{1-p(x_i;\beta)\}. \tag{0.1}$$

Generate $X = (1, X_1, X_2)$, where $X_j \sim N(0, I_N)$.

Set true parameter $\beta = (0.5, 1.2, -1)^\top$.

Set $N = 200, 500, 800, 1000$.

Estimate β using NR algorithm for $R = 200$ rounds of simulation. For each round of simulation, terminate the iteration when $\max_j |\widehat{\beta}_j^{old} - \widehat{\beta}_j^{new}| < 10^{-5}$. Denote $\widehat{\beta}_j^{(r)}$ as the estimation of β_j in the rth round of simulation. Then please: for each j, draw $(\widehat{\beta}_j^{(r)} - \beta_j)$ in boxplot for $N = 200, 500, 800, 1000$.

[proof]: $\frac{\partial l(\beta)}{\partial\beta} = \sum_{i=1}^{N} y_i x_i - \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)} \cdot x_i$

$= \sum_{i=1}^{N} (y_i - p(x_i;\beta)) x_i$

$\frac{\partial l^2}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{N} \partial\left(\left(1 - \frac{1}{1+\exp(x_i^T\beta)}\right)\cdot x_i\right)/\partial\beta^T$

$= \sum_{i=1}^{N} -\frac{\exp(x_i^T\beta)}{(1+\exp(x_i^T\beta))^2} x_i x_i^T$

$= -\sum_{i=1}^{N} p(x_i;\beta)(1-p(x_i;\beta))\, x_i \cdot x_i^T = -X^T W X$

Code is in "Problem2.html"

(3) 假设有m^+个正例和m^-个负例，令D^+与D^-分别表示正例、反例集合。定义排序“损失”如下：

$$\ell_{rank} = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(I(f(x^+)<f(x^-))+\frac{1}{2}I(f(x^+)=f(x^-))\right) \tag{0.2}$$

理解：若正例的预测值小于反例，则记一个“罚分”，若相等，则记0.5个罚分。 定义AUC：

$$AUC = 1-\ell_{rank}. \tag{0.3}$$

考虑一种简单的情况，即当数据中不存在$f(x^+)=f(x^-)$时，定义排序“损失”如下：

$$\ell_{rank} = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(I(f(x^+)<f(x^-))\right) \tag{0.4}$$

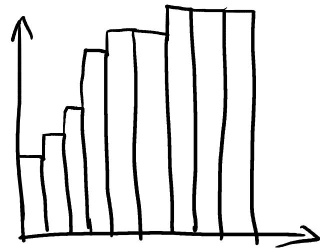
试证明以上定义的AUC即有限样本下ROC曲线下方的面积。

证明: $m^+ = TP+FN,\ m^- = TN+FP$

设前k个预测值的负例的集合为 D_k^-

则 $D_1^- \subseteq D_2^- \subseteq \cdots \subseteq D_{m^-}^-$ 且 $\biguplus_{k=1}^{m^-} D_k^- \setminus D_{k-1}^- = D^-$ 令 $D_k^- \setminus D_{k-1}^- = E_k$

对于离散ROC曲线 $\sum_{x^+\in D^+}\sum_{x^-\in E_k} I(f(x^+)>f(x^-))$ 为第k个柱体的面积



$$AUC = \frac{1}{m^+m^-}\sum_{k=1}^{m^-}\sum_{x^+\in D^+}\sum_{x^-\in E_k} I(f(x^+)>f(x^-))$$

$$= \frac{1}{m^+m^-}\sum_{k=1}^{m^-}\sum_{x^-\in E_k}\sum_{x^+\in D^+} I(f(x^+)>f(x^-))$$

$$= \frac{1}{m^+m^-}\sum_{x^-\in \biguplus_{k=1}^{m^-} E_k}\sum_{x^+\in D^+} I(f(x^+)>f(x^-))$$

$$= \frac{1}{m^+m^-}\sum_{x^-\in D^-}\sum_{x^+\in D^+} I(f(x^+)>f(x^-))$$

$\because I(f(x^+)>f(x^-)) + I(f(x^+)<f(x^-)) = 1$

$\therefore AUC = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}[1-I(f(x^+)<f(x^-))]$

$$= 1-\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-} I(f(x^+)<f(x^-)) = 1-L_{rank}$$