Q1.1 & 1.2)

Q4 1.1 We have showed earlier that $\frac{d\mu_i}{d\beta} = \mu_i(1-\mu_i)x_i$

Also $\frac{d\|\beta\|_2^2}{d\beta} = \frac{d}{d\beta}\left((\sqrt{\beta_1^2+\dots+\beta_d^2})^2\right) = \frac{d}{d\beta}(\beta_1^2+\dots+\beta_d^2) = \nabla_\beta(\beta_1^2+\dots+\beta_d^2)$

$= [2\beta_1 \dots 2\beta_d] = 2[\beta_1 \dots \beta_d]$

$\therefore \nabla_\beta l = 2\beta'\lambda - \sum_{i=1}^{n} \frac{y_i}{\mu_i}\mu_i(1-\mu_i)x_i + (1-y_i)\frac{}{1-\mu_i}\mu_i(1-\mu_i)x_i$

$= 2\beta'\lambda - \sum_{i=1}^{n} y_i x_i - y_i\mu_i x_i - \mu_i x_i + y_i\mu_i x_i$

$= 2\beta'\lambda - \sum_{i=1}^{n}(y_i-\mu_i)x_i$

$= 2\beta'\lambda - X^T(y-\mu)$ $\quad \beta' = \begin{bmatrix}0\\ \beta_1\\ \vdots\\ \beta_d\end{bmatrix}$

Q2. Let $g_j = 2\beta'\lambda - \sum_{i=1}^{n}(y_i-\mu_i)x_{ij}$

$\frac{\partial g_j}{\partial \beta_k} = 2\lambda - \sum_{i=1}^{n}(-\mu_i(1-\mu_i))x_{ik}x_{ij}$ since $y_i x_{ij}$ is const.

$H_{kj} = 2\lambda + \sum_{i=1}^{n}\mu_i(1-\mu_i)x_{ki}^T x_{ij}$

$= 2\lambda I' + X^T\begin{bmatrix}\mu_1(1-\mu_1) & & \\ & \ddots & \\ & & \mu_n(1-\mu_n)\end{bmatrix}X$ where $I' = \begin{bmatrix}0 & & \\ & \ddots & \\ & & 1\end{bmatrix}$

$= 2\lambda I' + X^T W X$

Q1.3)

$$\beta_{n+1} = \beta_n - (2\lambda I' + X^T W X)^{-1}(2\beta_n'\lambda - X^T(y-\mu))$$

Q1.4)

$mu^{(0)}$ = [0.9526 0.7311 0.7311 0.2689]

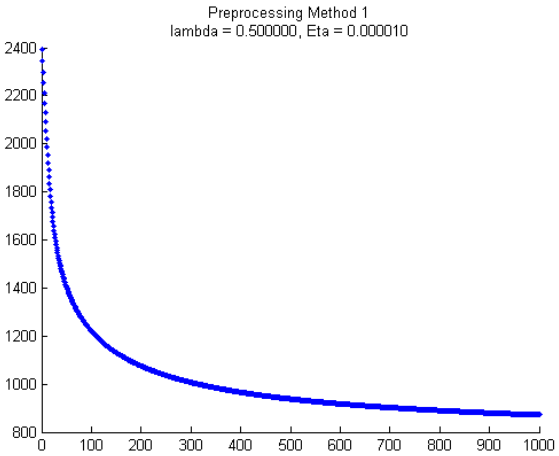
$beta^{(1)}$ = [-0.7479 1.4765 -2.2052]

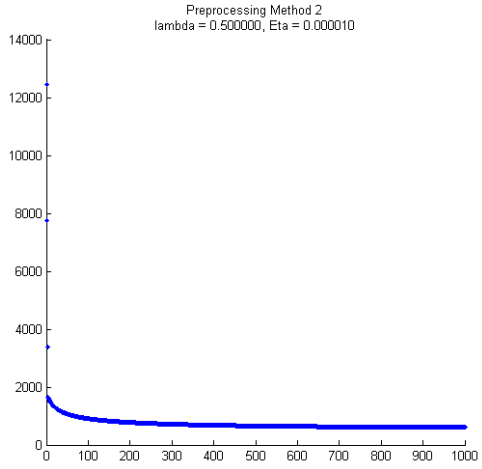
$mu^{(1)}$ = [0.9024 0.8140 0.3255 0.1860]

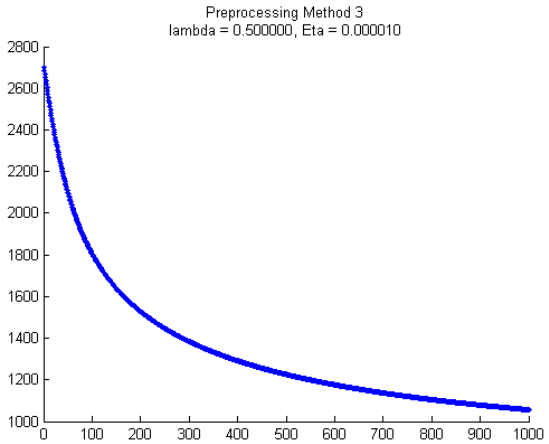
$beta^{(2)}$ = [-0.9729 1.5249 -2.0769]

Q2.1)

$$\beta_{n+1} = \beta_n - \delta(2\lambda\beta_n - X^T(y-\mu))$$

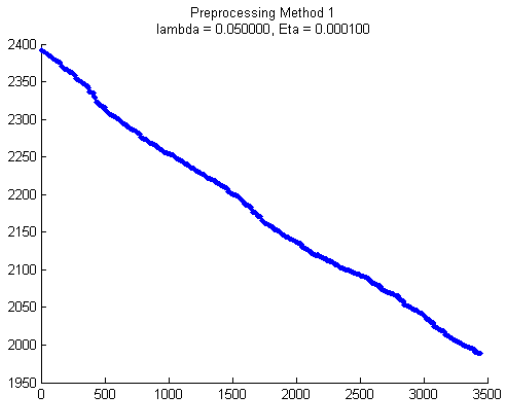


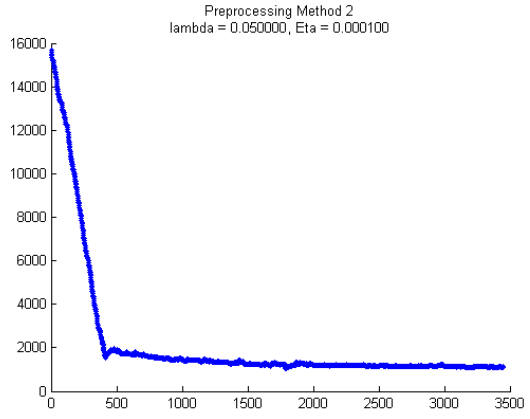


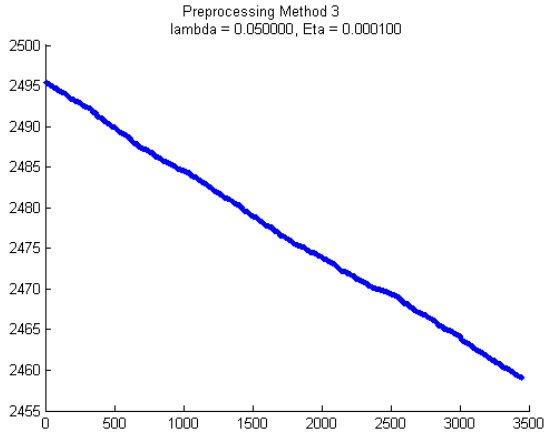


Q2.2)

$$\beta_{n+1} = \beta_n + \rho(2\beta' x - (y_i - \mu_i)x_i)$$







Q2.3)
They are in general better, but doesn't lead to the best cross-validation result.

Q2.4)
We have run through many sets of parameters of different orders of 10 by using a script we have written for automation. We have then filtered out the parameters which updated the beta values to perform far away from the minimum negative likelihood. After filtering these parameters, we are left with a smaller sets of parameters which we are capable to run by a script overnight. With 10-fold cross validation, we have figured out the parameters that performed the best, and it seems to be working well when we upload our result to Kaggle.

Best Parameters Found:
Lambda - 2
Eta - 0.00001
Preprocessing using transformation log(x_ij + 0.1)
Batch gradient descent