Homework II Yuting Sun.

P1

a. Proof:.

Given,
$$\sigma(t) = \frac{e^t}{1+e^t} \qquad 1-\sigma(t) = \frac{1}{1+e^t}$$

$$P(y_i=1|\vec{\theta}) = \sigma(\vec{\theta}^T\vec{x}_i)$$

$$\because P(y_i=1|\vec{\theta}) + P(y_i=0|\vec{\theta}) = 1$$

$$\therefore P(y_i=0|\vec{\theta}) = 1-P(y_i=0|\vec{\theta}) = 1-\sigma(\vec{\theta}^T\vec{x}_i)$$

$$P(y|\vec{\theta}) = P(y_1 \cdots y_n|\vec{\theta}) = \prod_{i=1}^{n}\left(P(y_i=1|\vec{\theta})^{y_i} \cdot P(y_i=0|\vec{\theta})^{1-y_i}\right)$$

$$= \prod_{i=1}^{n}\sigma(\vec{\theta}^T\vec{x}_i)^{y_i}(1-\sigma(\vec{\theta}^T\vec{x}_i))^{1-y_i}$$

$$= \prod_{i=1}^{n}\sigma(\vec{\theta}^T\vec{x}_i)^{y_i}\frac{1-\sigma(\vec{\theta}^T\vec{x}_i)}{(1-\sigma(\vec{\theta}^T\vec{x}_i))^{y_i}}$$

$$= \prod_{i=1}^{n}\left(\frac{\sigma(\vec{\theta}^T\vec{x}_i)}{1-\sigma(\vec{\theta}^T\vec{x}_i)}\right)^{y_i}(1-\sigma(\vec{\theta}^T\vec{x}_i))$$

$$\because \frac{\sigma(\vec{\theta}^T\vec{x}_i)}{1-\sigma(\vec{\theta}^T\vec{x}_i)} = \frac{e^t/\cancel{(1+e^t)}}{1/\cancel{(1+e^t)}} = e^t \quad \text{when } t=\vec{\theta}^T\vec{x}_i$$
$$= \exp(\vec{\theta}^T\vec{x}_i)$$

$$\therefore L(\vec{\theta}|y) = -\log P(y|\vec{\theta}) = -\sum_{i=1}^{n}\left(\log(e^{\vec{\theta}^T\vec{x}_i})^{y_i} - \log(1+e^{\vec{\theta}^T\vec{x}_i})\right)$$

$$= \sum_{i=1}^{n}\left[-y_i\vec{\theta}^T\vec{x}_i + \log(1+e^{\vec{\theta}^T\vec{x}_i})\right]$$

---

b. Proof
$$\nabla_\theta \log(\vec{\theta}|\vec{y}) = \sum_{i=1}^{n} -\vec{x}_i y_i + \underbrace{\frac{e^{\vec{\theta}^T\vec{x}_i}}{1+e^{\vec{\theta}^T\vec{x}_i}}}_{\text{scalar}}\vec{x}_i = -\sum_{i=1}^{n}\vec{x}_i(y_i-\sigma(\vec{\theta}^T\vec{x}_i))$$

$$\sigma(\vec{\theta}^T\vec{x}_i)$$

$$X = (\vec{x}_1^T, \vec{x}_2^T \cdots \vec{x}_n^T)^T$$

$$\sigma_\theta = (\sigma_1(\theta), \cdots \sigma_i(\theta) \cdots \sigma_n(\theta))^T$$

$$\sigma(\theta^T x_i) = \sigma_i(\theta)$$

$$\nabla(\vec{\theta}|y) = -X^T\vec{y} + X^T\sigma(\theta) = -X^T(\vec{y}-\sigma(\theta))$$

c).

$$\nabla_\theta \log(\vec{\theta}|\vec{y}) = -\sum_{i=1}^{n} \vec{x}_i(\underbrace{y_i}_{\text{Const}} - \sigma(\vec{\theta}^T\vec{x}_i))$$

$$\nabla_\theta^2 \log L(\vec{\theta}|\vec{y}) = +\sum_{i=1}^{n} \nabla_\theta (\vec{x}_i \cdot \sigma(\vec{\theta}^T\vec{x}_i))$$

$$\sigma_i(\theta) = \sigma(\vec{\theta}^T x_i) \qquad \nabla \cdot \vec{x}_i\, \sigma_i = \vec{x}_i \nabla\left(\frac{e^{\sigma(\vec{\theta}^T\vec{x}_i)}}{1+e^{\sigma(\vec{\theta}^T\vec{x}_i)}}\right) \cdot \frac{d_{\theta^T x}}{d\vec{\theta}}$$

$$= \vec{x}_i \cdot \frac{e^{\sigma(\theta)}(1+e^{\sigma(\theta)}) - e^{\sigma(\theta)} \cdot e^{\sigma(\theta)}}{(1+e^{\sigma_i(\theta)})^2} \cdot \vec{x}_i^T$$

$$= \vec{x}_i \frac{e^{\sigma_i(\theta)}}{(1+e^{\sigma_i(\theta)})^2} \cdot \vec{x}_i^T$$

$$= \vec{x}_i \frac{e^{\sigma_i(\theta)}}{1+e^{\sigma_i(\theta)}} \frac{1}{1+e^{\sigma_i(\theta)}} \vec{x}_i^T$$

$$= \vec{x}_i \underbrace{\sigma_i(\theta)(1-\sigma_i(\theta))}_{\text{scalar}} \vec{x}_i^T = \vec{x}_i \vec{x}_i^T \sigma_i(\theta)(1-\sigma_i(\theta))$$

$$\Longrightarrow \nabla_\theta^2 \log L(\vec{\theta}|\vec{y}) = \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \sigma_i(\theta)(1-\sigma_i(\theta))$$

$$= X^T D(\theta) X$$

$$D_{ii}(\theta) = \sigma_i(\vec{\theta}^T\vec{x}_i)(1-\sigma_i(\vec{\theta}^T\vec{x}_i)) = \sigma_i(\theta)(1-\sigma_i(\theta))$$

diagonal matrix.

d) Proof.

X is full rank. $-\nabla_\theta^2 \log L(\theta|y) = + \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \sigma_i(\vec{\theta}^T \vec{x}_i)(1-\sigma(\vec{\theta}^T \vec{x}_i))$

Pick up a vector $\underset{p\times 1}{\vec{y}}$ $(\vec{y} \neq \vec{0})$

$$\vec{y}^T \nabla_\theta^2 (\log L(\theta|y)) \vec{y}$$

$$= \vec{y}^T \left(\sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \frac{e^{\vec{\theta}^T \vec{x}_i}}{(1+e^{\vec{\theta}^T \vec{x}_i})^2} \right) \vec{y}$$

$$= \sum_{i=1}^{n} (\vec{y}^T \vec{x}_i)^2 \frac{e^{\vec{\theta}^T \vec{x}_i}}{(1+e^{\vec{\theta}^T \vec{x}_i})^2}.$$

$\because e^{\vec{\theta}^T \vec{x}_i} > 0 \implies \frac{e^{\vec{\theta}^T \vec{x}_i}}{1+e^{\vec{\theta}^T \vec{x}_i}} > 0$.

$\because$ X full rank. $\therefore (\vec{y}^T \vec{x}_i)^2 = 0$ only when $\vec{y} = \vec{0}$

However, $\vec{y} \neq 0$, thus $(\vec{y}^T \vec{x}_i)^2 > 0$

Therefore, if X is full rank, the Hessian is positive definite (PD)

If $N(X)$ is non-trivial.

$$\vec{y}^T \nabla_\theta^2 (\log L(\theta|y)) \vec{y}$$

$$= \sum_{i=1}^{n} (\vec{y}^T \vec{x}_i)^2 \frac{\vec{\theta}^T \vec{x}_i}{(1+e^{\vec{\theta}^T \vec{x}_i})^2}.$$

$\because N(X) \neq 0$.

By picking suitable $\vec{y} \neq \vec{0}$, it is possible that $(\vec{y}^T \vec{x}_i)^2 = 0$.

$$\implies \vec{y}^T \nabla_\theta^2 (\log L(\theta|y)) \vec{y} \geq 0$$

Therefore, if $N(X)$ is non-trivial, the Hessian is positive semi-definite (PSD)

e) $\theta^{(l+1)} = \theta^{(l)} - (\nabla_\theta^2 \log L(\theta|y)|_{\theta=\theta^{(l)}})^{-1} \nabla_\theta \log L(\theta|y)|_{\theta=\theta^{(l)}}$

The Newton-Raphson step is:

$$\theta^{(l+1)} = \theta^{(l)} - (X^T D(\theta^{(l)}) X)^{-1} (-X^T(y - 6(\theta^{(l)})))$$

$$= \theta^{(l)} + (X^T D(\theta^{(l)}) X)^{-1} X^T (y - 6(\theta^{(l)}))$$

$$= (X^T D(\theta^{(l)}) X)^{-1} \cdot X^T D(\theta) X \theta^{(l)} + (X^T D(\theta^{(l)}) X)^{-1} X^T (y - 6(\theta^{(l)}))$$

$$= (X^T D(\theta) X)^{-1} (X^T D(\theta) X \theta^{(l)} + X^T D(\theta) \cdot D^{-1}(\theta^{(l)}) (y - 6(\theta^{(l)})))$$

$$= (X^T D(\theta^{(l)}) X)^{-1} X^T D(\theta^{(l)}) \underbrace{(X\theta^{(l)} + D^{-1}(\theta^{(l)}) (y - 6(\theta^{(l)})))}_{Z \text{ (response)}}$$

$$= (X^T D(\theta^{(l)}) X)^{-1} X^T D(\theta^{(l)}) Z \quad ①$$

Compared with least-square model,

$$\hat{\theta}_{LS} = \arg\min \tfrac{1}{2} (\vec{y} - X\vec{\theta}) W (\vec{y} - X\vec{\theta})^T$$

$$\nabla \hat{\theta}_{LS} = 0. \Rightarrow \vec{\theta} = (\vec{X}^T W. X)^{-1} X^T W \vec{y}. \quad ②$$

According to ①② $\quad Z \underset{\text{equivalent}}{\Longleftrightarrow} \vec{y}$.

therefore. $\theta^{(l+1)} = \arg\min_\theta \frac{1}{2} (\tilde{y}(\theta^{(l)}) - X\theta)^T W(\theta^{(l)}) (\tilde{y}(\theta^{(l)}) - X\theta)$

$$\left(\begin{array}{l} W(\theta^{(l)}) = D(\theta^{(l)}) \\ \tilde{y}(\theta^{(l)}) = X\theta^{(l)} + D^{-1}(\theta^{(l)})(y - 6(\theta^{(l)})) \end{array} \right)$$

P2.

Proof. $f: R^n \to R$ is a convex function. $\Rightarrow \nabla^2 f \succeq 0$

$A\vec{x} + \vec{b} \in \mathrm{dom} f$.

$$\underbrace{\bar{\lambda} = 1-\lambda}$$

f convex $\Rightarrow f(\lambda\vec{x} + \bar{\lambda}\vec{y}) \le \lambda f(\vec{x}) + \bar{\lambda} f(\vec{y})$

(To prove $g(\vec{x}) = f(A\vec{x} + \vec{b})$ convex, we hope. $g(\lambda\vec{x} + \bar{\lambda}\vec{y}) \le \lambda g(\vec{x}) + \bar{\lambda} g(\vec{y})$.)

$g(\vec{x}) = f(A\vec{x} + \vec{b})$

$$g(\lambda\vec{x} + \bar{\lambda}\vec{y}) = f(A(\lambda\vec{x} + \bar{\lambda}\vec{y}) + \vec{b})$$

$\underbrace{\lambda + \bar{\lambda} = 1}$
$$= f(A(\lambda\vec{x} + \bar{\lambda}\vec{y}) + (\lambda + \bar{\lambda})\vec{b})$$

$$= \underbrace{f(\lambda(A\vec{x} + \vec{b}) + \bar{\lambda}(A\vec{y} + \vec{b}))} \quad ①$$

$f(\vec{x})$ is convex, $\Rightarrow ① \le \lambda f(A\vec{x} + \vec{b}) + \bar{\lambda}(A\vec{y} + \vec{b})$

$$= \lambda g(\vec{x}) + \bar{\lambda} g(\vec{y})$$

Therefore. $g(\vec{x})$ is also a convex function.

P3. Given:

X_* is a local minimizer of a convex function,

assume. X_* could not be a global minimizer, We can find a point $y \in \mathbb{R}$ such that $f(y) < f(x^*)$. $X_*, y \in \Omega$

As a local minimizer, $\exists \epsilon > 0$ s.t $f(X_*) \leq f(x)$ for $\forall x$ s.t $\|x - X_*\|_2 < \varepsilon$. We now consider a point x in the line segement between x_* and y.

$$X = \bar{\lambda} x_* + \lambda y \qquad (\bar{\lambda} = 1-\lambda) \qquad \lambda \in (0,1) \quad X \in \Omega.$$

By convexity, $\underbrace{f(x) = f(\lambda y + \bar{\lambda} x_*)}_{①} \leq \lambda f(y) + \bar{\lambda} f(x_*)$

$f(y) < f(x_*)$ $\qquad f(x) \leq \lambda f(X_*) + (1-\lambda) f(X_*) = f(X_*)$

Since there exists a point. $f(X_*) \leq f(\lambda y + (1-\lambda) X_*)$ ②

① ② $\quad f(x) = f(\lambda y + \bar{\lambda} X_*) \leq f(X_*) \leq f(\lambda y + \bar{\lambda} X_*)$

It is not possible, thus X_* is not a local minimizer any more, and should be a global minimizer,

---

Assume the minimizer is not unique $\Rightarrow$ there exits two points $a, b \in \Omega$ such that $f(a)$, $f(b)$ are both local minima. $f(a) \geq f(b)$ WLoG

$\lambda \in (0,1)$, $1-\lambda > 0$ $\quad \lambda f(a) + (1-\lambda) f(b) \leq \lambda f(a) + (1-\lambda) f(a) = f(a)$

f - strictly convex $\Rightarrow f(\lambda a + (1-\lambda) b) < \lambda f(a) + (1-\lambda) f(b)$

Since: $x = \lambda a + (1-\lambda) b \in \Omega \Rightarrow f(x) < f(a)$ $\qquad < \lambda f(a) + (1-\lambda) f(a) = f(a)$

$f(a)$ is not a minimizer, which contradicts out assumption.

Therefore, if the function is strictly convex, there is a unique minimizer.

P4. Given.

$g_i(x|\bar{x})$ is a majorizer of $f_i(x)$,

we know. $g_i(x|\bar{x}) \geq f_i(x)$ for $\forall x$.

$g_i(\bar{x}|\bar{x}) = f_i(\bar{x})$

When $g(x|\bar{x}) = \sum_{i=1}^{n} g_i(x|\bar{x})$

$= g_1(x|\bar{x}) + g_2(x|\bar{x}) + \cdots g_i(x|\bar{x}) + \cdots + g_n(x|\bar{x})$

(each term is greater than. $f_i(x)$ $i = 1, 2 \cdots n$)

$\geq f_1(x) + f_2(x) + \cdots f_i(x) + \cdots + f_n(x)$

$= \sum_{i=1}^{n} f_i(x)$ ①

$g(\bar{x}|\bar{x}) = \sum_{i=1}^{n} g_i(\bar{x}|\bar{x})$

$= g_1(\bar{x}|\bar{x}) + \cdots g_i(\bar{x}|\bar{x}) + \cdots + g_n(\bar{x}|\bar{x})$

(each term is equal to. $f_i(\bar{x})$ $i = 1, 2 \cdots n$)

$= f_1(\bar{x}) + \cdots f_i(\bar{x}) + \cdots + f_n(\bar{x})$

$= \sum_{i=1}^{n} f_i(\bar{x})$ ②

Based on ①②. $g(x|\bar{x}) = \sum_{i=1}^{n} g(x|\bar{x})$ is a majorizer of $f(x) = \sum_{i=1}^{n} f_i(x)$.

P5.

a) $l_i(\theta) = -y_i\theta^T x_i + \log(1+e^{\theta^T x_i})$

$$g_i(\bar\theta|\bar\theta) = l_i(\bar\theta) \underbrace{- \left(y_i - \frac{e^{\bar\theta^T x_i}}{1+e^{\bar\theta^T x_i}}\right) x_i^T(\bar\theta-\bar\theta)}_{0} + \underbrace{\frac{1}{8}\|\bar\theta-\bar\theta\|_2^2}_{0}$$

$$= l_i(\bar\theta) \quad ①$$

According to Taylor's theorem,

$$f(\theta) \le f(\bar\theta) + \nabla f(\bar\theta)^T(\theta-\bar\theta) + \frac{1}{2}(\theta-\bar\theta)^T M(\theta)(\theta-\bar\theta)$$

(M is positive definite matrix)

$$\nabla l_i(\theta) = -(y_i - \sigma(\theta^T x_i))x_i^T$$

$$l_i(\vec\theta) \le l_i(\bar\theta) - (y_i-\sigma(\theta^T x_i))x_i^T(\theta-\bar\theta) + \frac{1}{2}(\theta-\bar\theta) M(\theta-\bar\theta)$$

Since, $g_i(\theta|\bar\theta) = l_i(\bar\theta) - (y_i-\sigma(\bar\theta^T x_i))x_i^T(\theta-\bar\theta) + \frac{1}{8}(\theta-\bar\theta)^T x_i x_i^T(\theta-\bar\theta)$

We need to compare $\frac{1}{2}(\theta-\bar\theta)^T M(\theta-\bar\theta)$ with $\frac{1}{8}(\theta-\bar\theta)^T x_i x_i^T(\theta-\bar\theta)$

$$M = \nabla^2 l_i(\theta) = \vec{x}_i(\sigma(\theta^T\vec{x}_i)(1-\sigma(\theta^T\vec{x}_i)))\vec{x}_i^T$$

$$\sigma(\theta)(1-\sigma(\theta)) \xRightarrow[t=\theta^T\vec{x}_i]{} \frac{e^t}{(1+e^t)^2}$$

$$\frac{e^t}{(1+e^t)^2} - \frac{1}{4} = \frac{4e^t-(1+e^t)^2}{4(1+e^t)^2} = \frac{-(1-e^t)^2}{4(1+e^t)^2} \le 0$$

$$\Rightarrow \frac{e^t}{(1+e^t)^2} \le \frac{1}{4} \qquad MI \le \frac{1}{4}\vec{x}_i\cdot\vec{x}_i^T \Rightarrow \underline{\frac{1}{2}MI \le \frac{1}{8}\vec{x}_i\vec{x}_i^T}$$

$$\Rightarrow \frac{1}{2}(\theta-\bar\theta)^T M(\theta-\bar\theta) \le \frac{1}{8}(\theta-\bar\theta)^T x_i x_i^T(\theta-\bar\theta)$$

$$g_i(\theta|\bar\theta) \ge l_i(\vec\theta) \quad ②$$

Based on ①② $g(\theta|\bar\theta)$ is a majorizer of $l_i(\theta)$

b) To minimize. $g_i(\theta|\bar\theta) = l_i(\bar\theta) - (y_i - 6(\bar\theta^T x_i)) x_i^T(\theta-\bar\theta)$
$+ \frac{1}{8}(\theta-\bar\theta)^T x_i x_i^T(\theta-\bar\theta)$

equivalent to minimize

$$-\underbrace{(y_i - 6(\bar\theta^T x_i))}_{\alpha} x_i^T(\theta-\bar\theta) + \frac{1}{8}(\theta-\bar\theta)^T x_i x_i^T(\theta-\bar\theta)$$

Construct a quadratic function $\Longrightarrow$

$$-\alpha x_i^T(\theta-\bar\theta) + \left(\frac{1}{2\sqrt2}\right)^2(\theta-\bar\theta)^T x_i x_i^T(\theta-\bar\theta) + (\sqrt2)^2\alpha^2$$

$$= \left(\frac{1}{2\sqrt2}(\theta-\bar\theta)^T x_i - \sqrt2\alpha\right)^2 = \frac{1}{8}\left((\theta-\underset{\text{const}}{\bar\theta})^T x_i - 4\alpha\right)^2$$

For the sum of all minimized terms,

$$\min_\theta \sum_{i=1}^n g_i(\theta|\bar\theta) \Longleftrightarrow \min_\theta \sum_{i=1}^n \frac{1}{2}\cdot\frac{1}{4}\left(\theta^T x_i - (\bar\theta^T x_i + 2\alpha)\right)^2 \cdot \quad (\alpha = y_i - 6(\bar\theta^T x_i))$$

$$= \frac{1}{2}\cdot(\tilde y(\bar\theta) - X\theta)^T W(\tilde y(\bar\theta) - X\theta)$$

$$\tilde y(\bar\theta) = X\bar\theta + 4(y - 6(\bar\theta^T X))$$

$$W = \frac{1}{4}I.$$

c). According to Taylor's Theorem,

f is a convex function and $\underline{\nabla^2 f(\theta) \le \mu I}$.
PSD

$$f(\theta) \le f(\bar\theta) + \nabla f^T(\bar\theta)(\theta-\bar\theta) + \frac{1}{2}(\theta-\bar\theta)^T\mu I(\theta-\bar\theta)$$

$$= f(\bar\theta) + \nabla f(\bar\theta)(\theta-\bar\theta) + \frac{1}{2}\mu\|\theta-\bar\theta\|_2^2 = l(\theta) \quad ①$$

On the other hand

$$l(\bar\theta) = f(\bar\theta) + \nabla f^T(\bar\theta)\underbrace{(\bar\theta-\bar\theta)}_{0} + \frac{1}{2}(\bar\theta-\bar\theta)^T\mu I\underbrace{(\bar\theta-\bar\theta)}_{0} \quad ②$$

Based on ①②. $l(\theta)$ is a majorizer of $f(\theta)$.

P6

a). $\min. \sum_{i=1}^{n} -y_i \beta^T x_i + \log(1+e^{\beta^T x_i})$

$N(X) = \vec{r} = [1\ 1\ 1 \ldots 0]^T$

If we pick a vector. $\vec{\beta}' = \vec{\beta} + \vec{r}t$ (t - const ($\neq 0$))

$\min \sum_{i=1}^{n} -y_i \beta'^T x_i + \log(1+e^{\beta'^T x_i})$

$= \min \sum_{i=1}^{n} -y_i (\vec{\beta} + \vec{r}t)^T \vec{x}_i + \log(1+e^{(\beta^T + \vec{r}t)^T \vec{x}_i})$

($\because \vec{r} \cdot t\, x_i = 0$)

$= \min \sum_{i=1}^{n} -y_i (\vec{\beta})^T \vec{x}_i + \log(1+e^{\vec{\beta}^T x_i})$

Therefore. $\vec{\beta}' = \vec{\beta} + \vec{r}t$ ($\vec{\beta}' \neq \vec{\beta}$) can also minimize the function $\Rightarrow$ the solution is not unique.

b) $\beta^{(\ell+1)} = \arg\min \frac{1}{2} (\tilde{y}(\beta^{(\ell)}) - X\beta)^T W(\beta^{(\ell)}) (\tilde{y}(\beta^{(\ell)}) - X\beta)$

$$\left(\begin{aligned} \tilde{y}(\beta^{(\ell)}) &= X\beta^{(\ell)} + D(\beta^{(\ell)})^{-1} (y - 6(\beta^{(\ell)})) \\ W(\beta^{(\ell)}) &= D(\beta^{(\ell)}) \\ D_{ii}(\beta^{(\ell)}) &= 6_i(\beta^{(\ell)})(1 - 6_i(\beta^{(\ell)})) \end{aligned} \right)$$

$$g^{(\ell+1)}(\beta) = \frac{1}{2} (\tilde{y}(\beta^{(\ell)}) - X\beta)^T W(\beta^{(\ell)}) (\tilde{y}(\beta^{(\ell)}) - X\beta) + \lambda r^T \beta$$

$$\nabla g(\beta) = -X^T W (\tilde{y}(\beta^{(\ell)}) - X\vec{\beta}) + \lambda \vec{r} = \vec{0}$$

$$\nabla g(\lambda) = \vec{r}^T \vec{\beta} = 0$$

$$\begin{bmatrix} X^T W X & \vec{r} \\ \vec{r}^T & 0 \end{bmatrix} \begin{bmatrix} \vec{\beta} \\ \lambda \end{bmatrix} = \begin{bmatrix} X^T W \vec{\tilde{y}} \\ 0 \end{bmatrix}$$

$$L(\beta) = \frac{1}{2} (\tilde{y}(\beta^{(\ell)}) - X\beta)^T W(\beta^{\ell}) (\tilde{y}(\beta^{\ell}) - X\beta) + \lambda r^T \beta$$

(see code)

c). Similar to b), only need to change $W(\beta^{\ell}) = \frac{1}{4} I$,

$$\tilde{y}(\beta^{\ell}) = X\beta^{(\ell)} + 4(y - 6(\vec{\theta}^T X))$$

(see code)

d). see code

P6 (see code)
Comments for part e:

Compared with the win/loss percentage, the ranking estimation using logistic regression basically has the similar results but not exactly the same. As seen from the ranking results, some teams' betas are very close, which leads to a different order among 2 or 3 teams from the win/loss ranking. This makes sense because the score of each team should be independent but our derived non-linear model takes the score differences (y = 1,0) between 2 teams into consideration. The underlying assumption is that these 30 teams may have correlations, leading to the differences between our results and the win/loss ranking.

Compared with the results from linear model we applied in the last homework, logistic regression seems to have a better performance. (Logistic regression predicts 12 correct rankings but linear regression predicts 11 correct rankings). It seems that for this NBA ranking, logistic regression performs well since score differences are converted into (0,1) binary number, which makes the classification more robust to numerical factors.