Q1 $b = -34.084$

$$w = \begin{bmatrix} 0.6487 \\ -0.0632 \end{bmatrix}$$

Test error = 21.4859

Predicted response at $x = [100, 100]$ = 24.4581

Q2 $$J = \min_{w,b} \sum_{i=1}^{n} c_i (y_i - w^T x_i - b)^2$$

$$\min_{w,b} \sum_{i=1}^{n} c_i (y_i - \theta^T x_i)^2$$

$$= \cancel{(Y}$$

$x_i = [1, \ldots\ldots x_p]$

$\theta = [b, w_1 \ldots\ldots w_p]^T$

$$= (Y - X\theta)^T C (Y - X\theta)$$

where $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $X = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(p+1)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \cdots & x_n^{(p+1)} \end{bmatrix}$ $C = \begin{bmatrix} c_1 & & & 0 \\ & c_2 & & \\ & & \ddots & \\ 0 & & & c_n \end{bmatrix}$

$$\theta = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_p \end{bmatrix}$$

$$J = (Y^T - \theta^T X^T)(CY - CX\theta)$$

$$J = \underbrace{Y^T C Y}_{0} - \underbrace{Y^T C X \theta - \theta^T X^T C Y}_{\text{same}} + \theta^T X^T C X \theta$$

$$\nabla J = -2X^T C Y + 2X^T C X \theta = 0$$

$$\therefore \quad \theta = (X^T C X)^{-1} X^T C Y$$

Q3(a) For ρ of robust regression, the MM algorithm has

$$\rho_t(r) = \underbrace{\rho(r_{t,i})}_{\substack{0 \\ \text{as constant} \\ \text{wrt } \theta}} - \underbrace{\frac{1}{2} r_{t,i} \psi(r_{t,i})}_{\substack{0 \\ \text{as constant} \\ \text{wrt } \theta}} + \underbrace{\frac{1}{2} \frac{\psi(r_{t,i})}{r_{t,i}}}_{C_{t,i}} \underbrace{r^2}_{\substack{(Y-X\theta)^T \\ (Y-X\theta)}}$$

~~$\frac{\partial \rho_t(r)}{\partial r} = C_{t,i}\, r$~~

$$\frac{\partial \rho_t(r)}{\partial \theta} = \frac{d}{\partial \theta} (Y - X\theta)^T C_t (Y - X\theta) = 0$$

where $C_t = \begin{bmatrix} C_{t,1} & & 0 \\ & \ddots & \\ 0 & & C_{t,n} \end{bmatrix}$

$\therefore \quad \underline{\theta} = (X^T C_t X)^{-1} X^T C Y$

Hence the same form as weighted linear reg. Also, the weights (ie $C_{t,i} \approx f(r_{t,i})$) are some function of the residuals. Therefore it is called iteratively reweighted least squares.

As, we can see above, ~~the~~ ~~weight~~ ~~o~~

$$w_i = \frac{\psi(r_{t,i})}{2\, r_{t,i}}$$

& $\psi(r_{t,i}) / r_{t,i}$ is non increasing for $r > 0$.

ie as $r \uparrow$ $w_i \downarrow$ or remains same.

ie More weight is given to inliers than outliers whereas in case of OLS,

$$w_i = 1$$

ie same weight for inliers as well as outliers.

& hence the algo achieves robustness.

Q4 a) $J_t(\theta) = J(\theta^{(t)}) + \nabla J(\theta^{(t)})^T(\theta - \theta^{(t)}) + \frac{1}{2}(\theta - \theta^{(t)})^T B(\theta - \theta^{(t)})$

Using Taylor series expansion,

$$J(\theta) = J(\theta^{t}) + \nabla J(\theta^{t})^T(\theta - \theta^{(t)}) + \frac{1}{2}(\theta - \theta^{(t)})^T \nabla^2 J(\tilde{\theta})(\theta - \theta^{(t)})$$

$$\therefore J_t(\theta) - J(\theta) = \frac{1}{2}(\theta - \theta^{(t)})^T(B - \nabla^2 J(\tilde{\theta}))(\theta - \theta^{(t)})$$

$\because \quad B - \nabla^2 J(\tilde{\theta})$ is PSD $\forall \tilde{\theta}$

$\therefore \quad x^T(B - \nabla^2(\tilde{\theta}))x \geq 0 \ \forall x$

Let $\theta = \theta - \theta^{(t)}$

Then,

$$(\theta - \theta^{(t)})^T(B - \nabla^2(\tilde{\theta}))(\theta - \theta^{(t)}) \geq 0$$

or $J_t(\theta) - J(\theta) \geq 0$

or $J_t(\theta)$ is majorizing function for J.

$$\min J_t(\theta) = \underbrace{\cancel{J(\theta^{(t)})}}_{0} + \nabla J(\theta^{(t)})^T \theta - \underbrace{\cancel{\nabla J(\theta^{(t)})^T \theta^{(t)}}}_{0}$$

$$+ \frac{1}{2}\left(\theta^T B \theta - \theta^T B \theta^{(t)} - \theta^{(t)T} B \theta + \underbrace{\cancel{(\theta^{(t)})^T B \theta^{(t)}}}_{0}\right)$$

$$\nabla J_t(\theta) = \nabla J(\theta^{(t)}) + \frac{1}{2}\left(2B\theta - 2B\theta^{(t)}\right)$$

$$0 = \nabla J(\theta^{(t)}) + B(\theta - \theta^{(t)})$$

$$\therefore \ \theta = \theta^{(t)} - B^{-1}(\nabla J(\theta^{(t)}))$$

b) $$J(\theta) = -l(\theta) + \lambda\|\theta\|^2$$

$$= \sum \log(1+\exp(-y_i \theta^T \tilde{x}_i)) + \lambda\|\theta\|^2$$

$$J(\theta) = \frac{1}{n}\sum \log(1+\exp(-y_i\theta^T\tilde{x}_i)) + \lambda\theta^T\theta$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} \underbrace{\frac{1}{1+\exp(-y_i\theta^T\tilde{x}_i)}}_{\sigma} \times -y_i\tilde{x}_i + 2\lambda\theta$$

$$\therefore \ \frac{\nabla J(\theta)}{\partial\theta} = \cancel{\sum(1-\sigma)(-y_i\tilde{x}_i) + 2\lambda\theta}$$

$$\frac{\nabla J(\theta)}{\partial\theta} = \sum \tilde{x}_i\left(\frac{e^{\theta^T\tilde{x}_i}}{1+e^{\theta^T\tilde{x}_i}} - y_i\right) + 2\lambda\theta$$

$$\& \ \frac{\nabla^2 J(\theta)}{\partial\theta^2} = \sum \tilde{x}_i\tilde{x}_i^T\left(\frac{e^{\theta^T\tilde{x}_i}}{(1+e^{\theta^T\tilde{x}_i})^2}\right) + 2\lambda I$$

$$\max\left(\frac{\nabla^2 J(\theta)}{\partial\theta^2}\right) = \max\left(\sum \tilde{x}_i\tilde{x}_i^T \underbrace{\{\sigma(1-\sigma)\}}_{\max = 1/4} + 2\lambda I\right)$$

$$\therefore \max\left(\frac{\nabla^2 J(\theta)}{\cancel{\partial\theta}}\right) = \sum \tilde{x}_i \tilde{x}_i^T \times \frac{1}{4} + 2\lambda I$$

$\rightarrow (p \times (p+1))$
$n \times p^n$

$$\max\left(\frac{\nabla^2 J(\theta)}{\cancel{\partial\theta}}\right) = \underbrace{\frac{\tilde{X}^T X}{4} + 2\lambda I}$$

Let this be B, then $B - \nabla^2 J(\theta)$ will always be PSD.

$$\& \quad \theta^{(t+1)} = \theta^{(t)} - \left(\frac{X^T X}{4} + 2\lambda I\right)^{-1} \nabla J(\theta^{(t)})$$

$$\theta^{(t+1)} = \theta^{(t)} - \left(\frac{(X^T X)^{-1}}{4} + 2\lambda I\right) \nabla J(\theta^{(t)})$$

where $\nabla J(\theta^{(t)}) = \sum \tilde{x}_i (1 - \sigma_i - y_i) + 2\lambda\theta$

$$\sigma_i = \frac{1}{1 + \exp(-y_i \theta^T \tilde{x}_i)}$$

~~On~~

The Computational advantage of mm algo compared to Newton's method:-

1) We have calculate just $X^T X$ instead of Hessian.

2) Inverse of a symmetric matrix is computationally less intensive than that of a non symmetric one. because ~~$(A^T)^{-1}$~~ its inverse will also be symmetric