Q1:

for 150cm:
distance to all samples

$$\begin{cases} P_1 = \sqrt{(150-171)^2} = 21 \\ P_2 = \sqrt{(150-168)^2} = 18 \\ P_3 = 41 \\ P_4 = 32 \\ P_5 = 0 \\ P_6 = 28 \end{cases}$$

for 155cm:
distance to all samples

$$\begin{cases} P_1 = \sqrt{(155-171)^2} = 16 \\ P_2 = 13 \\ P_3 = 36 \\ P_4 = 27 \\ P_5 = 5 \\ P_6 = 23 \end{cases}$$

for 165cm
distance to all samples

$$\begin{cases} P_1 = 6 \\ P_2 = 3 \\ P_3 = 26 \\ P_4 = 17 \\ P_5 = 15 \\ P_6 = 13 \end{cases}$$

for 190cm
distance to all samples

$$\begin{cases} P_1 = 19 \\ P_2 = 22 \\ P_3 = 1 \\ P_4 = 8 \\ P_5 = 40 \\ P_6 = 12 \end{cases}$$

When k=3

1. for height = 150 cm: nearest: P_1, P_2, P_5
2. for height = 155 cm: nearest: P_1, P_2, P_5
3. for height = 165 cm: nearest: P_1, P_2, P_6
4. for height = 190 cm: nearest: P_3, P_4, P_6

estimate height:

for height = 150 cm:

$$\hat{y}_{KNN} = \frac{80+78+65}{3} \approx 74.33\,kg$$

for height = 155 cm

$$\hat{y}_{KNN} = \frac{80+78+65}{3} \approx 74.33\,kg$$

for height = 165 cm

$$\hat{y}_{KNN} = \frac{80+78+83}{3} \approx 80.33\,kg$$

for height = 190 cm

$$\hat{y}_{KNN} = \frac{100+80+83}{3} \approx 87.67\,kg$$

Q2:

for height = 150cm

$\frac{1}{0} \approx 0$: $\hat{y}_{KNN} = \dfrac{(0 \cdot 65 + \frac{1}{18} \cdot 78 + \frac{1}{21} \cdot 80)}{(\frac{1}{18} + \frac{1}{21} + 0)} \approx 74.06\,kg$

for height = 155cm

$\hat{y}_{KNN} = \dfrac{(\frac{1}{16} \cdot 80 + \frac{1}{13} \cdot 78 + \frac{1}{5} \cdot 65)}{(\frac{1}{16} + \frac{1}{13} + \frac{1}{5})} \approx 70.71\,kg$

for height = 165cm

$\hat{y}_{KNN} = \dfrac{(\frac{1}{6} \cdot 80 + \frac{1}{3} \cdot 78 + \frac{1}{13} \cdot 83)}{\frac{1}{6} + \frac{1}{13} + \frac{1}{3}} \approx 79.24\,kg$

for height = 190cm

$\hat{y}_{KNN} = \dfrac{(\frac{1}{1} \cdot 100 + \frac{1}{8} \cdot 80 + \frac{1}{12} \cdot 83)}{1 + \frac{1}{8} + \frac{1}{12}} \approx 85.33\,kg$

I'm not sure if I should view $\frac{1}{0}$ as 0 or 1

$\frac{1}{0} \approx 1$ version:

$\hat{y}_{KNN} = \dfrac{(1 \cdot 65 + \frac{1}{18} \cdot 78 + \frac{1}{21} \cdot 80)}{1 + \frac{1}{18} + \frac{1}{21}} \approx 66.3\,kg$

Q3

$$J(x) = x^T Q x + d^T x + c$$

$$\nabla_x J(x) = \left(\frac{\partial J}{\partial x_1} + \frac{\partial J}{\partial x_2} + \frac{\partial J}{\partial x_3} \cdots \frac{\partial J}{\partial x_n}\right)^T$$

for $\frac{\partial J}{\partial x_i}$, $\frac{\partial J}{\partial x_i} = \underbrace{\left(\frac{\partial}{\partial x_i}\right)(x^T Q x)}_{①} + \underbrace{\left(\frac{\partial}{\partial x_i}\right)(d^T x)}_{②} + \underbrace{\left(\frac{\partial}{\partial x_i}\right) c}_{③}$

Seperate these 3 parts.

① : $(\partial/\partial x_i)(x^T Q x) = (\partial x^T/\partial x_i)(Qx) + x^T(\partial Qx/\partial x_i)$

since $Q = Q^T$

$$\partial Qx/\partial x_i = (\partial x^T Q/\partial x_i)^T, \ \frac{\partial Q^T}{\partial x_i} \cdot x = \frac{\partial Q}{\partial x_i} \cdot x$$

$\therefore \left(\frac{\partial}{\partial x_i}\right) x^T Q x = x^T Q + x^T \frac{\partial Q}{\partial x_i} x$

② : $\frac{\partial}{\partial x_i} d^T x = \left(\frac{\partial}{\partial x_i}\right)(d_1 x_1 + d_2 x_2 \cdots d_n x_n) = \left(\frac{\partial}{\partial x_i}\right)(d_i x_i) = d_i = d$

③ $\left(\frac{\partial}{\partial x_i}\right) c = 0$

$\therefore \frac{\partial J}{\partial x_i} = x^T Q + x^T\left(\frac{\partial Q}{\partial x_i}\right) x + d$

$\nabla_x J(x) = \left(x^T Q + x^T\left(\frac{\partial Q}{\partial x_1}\right) x + d, \ x^T Q + x^T \frac{\partial Q}{\partial x_2} x + d \ \cdots \ x^T Q + x^T\left(\frac{\partial Q}{\partial x_n}\right) x + d\right)^T$

to simplify the expression, we got $\nabla_x J(x) = 2Qx + d$

As for $H = \frac{\partial^2 J}{\partial x \partial x^T}$. $H_{ij} = \frac{\partial^2 J}{\partial x_i \partial x_j}$

$$\frac{\partial^2 J}{\partial x_i \partial x_j} = \left(\frac{\partial}{\partial x_i}\right)\left(\frac{\partial J}{\partial x_j}\right) = \left(\frac{\partial}{\partial x_i}\right)\left(\underset{①}{x^T Q} + \underset{②}{x^T\left(\frac{\partial Q}{\partial x_i}\right)x} + \underset{③}{d_i}\right)$$

for ①, since x is not depend on x_j. ∴ ① $= 0$

for ② $\left(\frac{\partial}{\partial x_j}\right)\left(x^T \frac{\partial Q}{\partial x_i} x\right) = \left(\frac{\partial}{\partial x_j}\right)\left(x_j \frac{\partial Q}{\partial x_i} x\right) = \frac{\partial Q}{\partial x_i} \cdot x_j$

for ③ d_i does not depend on x_j. ∴ ③ $= 0$

$$\therefore \frac{\partial^2 J}{\partial x_i \partial x_j} = \frac{\partial Q}{\partial x_i} \cdot x_j$$

due to $H = \frac{\partial^2 J}{\partial x \partial x^T}$

$$\therefore \quad H = \left(\frac{\partial Q}{\partial x_i}\right) \cdot x_j = 2Q$$

where each element $H_{ij} = \frac{\partial^2 J}{\partial x_i \partial x_j} = \left(\frac{\partial Q}{\partial x_i}\right) \cdot x_j$

$\therefore H = 2Q$

Q4:

$$\hat{y} = x'_{1\times p} * \beta$$

in this equation, $x'_{1\times p}$ is augmented by appending a column of 1s to account for the intercept term, resulting in a (p+1) dimension vector.

Since for KNN, we can view $\hat{y}$ as the weighted sum of the labels y of the k nearest neighbors in the training set, where weights are determined by β

When we set k=1, only to pick the nearest neighbor, the weights assigned to the features become equivalent to the coefficient β in linear regression.

Therefore, we can consider the linear regression as a special case of KNN with k=1.

Q5:

To show that $\hat{y} = X(X^TX)^{-1}X^Ty$ is a member of the column space of X, we need to show that $\hat{y}$ can be expressed as a linear combination of the columns of X

X^TX is a $(p+1)\times(p+1)$ symmetric matrix
When X^TX is invertible, it indicates that X^TX is full rank. So it is also the case of linear regression, that there is perfect multicollinearity among the features.

$X(X^TX)^{-1}$ is $n\times(p+1)$, and $X(X^TX)^{-1}X^T$ is $n\times n$ matrix
Therefore, y is a n-dimensional column vector, which leads to $\hat{y}$, the predicted value of dimension n.

consider $\hat{y}$ as linear combination of columns of X.

$$\hat{y} = X \cdot (X^TX)^{-1}X^Ty$$
$$= X \cdot [(X^TX)^{-1}X^Ty] = X \cdot \beta$$

this expression shows a linear combination of column X

Q6.

We need to prove that the dot product of difference vector and any column of X is 0

for $\hat{y} = \hat{\beta} X$

the difference vector $= y - \hat{y}$

to prove $y - \hat{y}$ is orthogonal to the column space of X. we can prove $y - \hat{y}$ with any column of X is 0

So consider a specific column of X is X_i where i ranges from 1 to $(p+1)$

$$(y - \hat{y}) \cdot X_i = (y - X\hat{\beta}) \cdot X_i = y X_i - (X\hat{\beta}) X_i$$
$$= y \cdot X_i - (X\hat{\beta})_1 X_{i1} - (X\hat{\beta})_2 X_{i2} \cdots - (X\hat{\beta})_n \cdot X_{in}$$

Since X_i represents the i^{th} column of X, the dot product $y \cdot X_i$ is the sum of the elements-wise product of y and x_i

elements of X_i ($X_{i1}, X_{i2} \cdots X_{in}$) are fixed. then the dot product of $(X\hat{\beta})_1 X_{i1} + (X\hat{\beta})_2 X_{i2} \cdots + (X\hat{\beta})_n X_{in}$ can be considered as linear combination of elements of $X\hat{\beta}$

$\therefore X\hat{\beta}$ is orthogonal to $y-\hat{y}$

so $(X\hat{\beta})_1 X_{i1} + (X\hat{\beta})_2 X_{i2} \cdots + (X\hat{\beta})_n X_{in} = 0$, which makes

$(y-\hat{y})X_i = 0$

Therefore, $(y-\hat{y})$ is orthogonal to the column space of X