4. 1. The wage equation is given as

$$W_i = \beta_0 + \beta_1 X_i + \mu_i$$

where W_i is the wage of worker i
X_i is the labour market experience of worker i

A)

We suspect that the intercept is different for men and women. To test this suspicion the model is changed as

$$W_i = \beta_0 + \beta_1 X_i + \beta_2 M + \mu_i$$

where M is a dummy variable that can take values 1 and 0. It takes value 1 for man and 0 for woman.
Now,

$W_i = \beta_0 + \beta_1 X_i + \beta_2 + \mu$, for man $M=1$

and

$= \beta_0 + \beta_1 X_i + \mu$, for woman $M=0$

If the value of β_2 is significant then our suspicion is correct and intercept is different for men and women. If β_2 is zero then our suspicion is incorrect.

B) We suspect that the slope is different for men and women. To test this suspicion the model is changed as

$$W_i = B_0 + (B_1 + B_2 M) X_i + u_i$$

where M is a dummy variable that can take values 1 and 0. It takes value 1 for man and 0 for woman.

Now,

$$W_i = B_0 + (B_1 + B_2) X_i + u_i, \text{ for man } M=1$$

and

$$= B_0 + B_1 X_i + u_i, \text{ for woman } M=0$$

If the value of B_2 is significant then our suspicion is correct and the slope is different for men and women. If B_2 is zero then our suspicion is incorrect.

c) We suspect that the relationship between wage and experience resembles an upward slope. To test this suspicion the model is changed as

$$W_i = \beta_0 + \beta_1 X_i^2 + \mu_i$$

We square the X_i term of the model to make it always positive.

Now if the value of W_i obtained is positive then it is an upward slope else it is downward.

Since the terms β_0 and μ_i can also be negative we can square them too, the updated equation becomes

$$W_i = \beta_0^2 + \beta_1 X_i^2 + \mu_i^2$$

Since All the other terms are positive the value obtained on RHS will help us determine if the slope is upward (positive)

2. L2 regularization or Ridge regression promotes smaller coefficients. It is used to prevent the model from overfitting. When we use a model with higher complexity the coefficients of the model become very large and the model starts to overfit and doesn't perform too well on the testing set. To prevent this we use ridge regression. in which a penalty term is added to the cost function

$$J(\theta) = \frac{1}{M} \sum_{i=1}^{M} \left(\theta^T \phi(x_i) - y_i \right)^2$$

$$+ \lambda \underbrace{\sum_{j=1}^{p} \theta_j^2}_{\text{penalty term.}}$$

Adding this term, reduces the value of coefficients and prevents overfitting

3. Suppose the parameters are given by $D = \{(x_i, y_i) : i \in [1, N]\}$

It is a linear model $y = \theta^T x + c$ and θ are the weights

Using Bayes Theorem we can write

$$P\left(\frac{\theta}{D}\right) = \frac{P\left(\frac{D}{\theta}\right) P(\theta)}{P(D)}$$

Taking log on both sides.

$$\log P\left(\frac{\theta}{D}\right) = \log\left(\frac{D}{\theta}\right) + \log P(\theta) - \log P(D)$$

Using maximum a-Posterior Inference the solution is given by

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \log\left(\frac{\theta}{D}\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \left(\log\left(\frac{D}{\theta}\right) + \log P(\theta) - \log P(D) \right)$$

Using MAP we can ignore $-\log P(D)$

$$= \underset{\theta}{\operatorname{argmax}} \left(\log\left(\frac{D}{\theta}\right) + \log P(\theta) \right) \quad -(1)$$

DELTA Pg No.

Log likelihood is given by

$$\log P\left(\frac{D}{\theta}\right) = \log P(Y \mid X, \theta)$$

$$= \log \prod_{n=1}^{N} P(y_n \mid x_n, \theta)$$

$$= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n - \theta^T x_n)^2}{2\sigma}}$$

$$= \sum_{n=1}^{N} \left(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_n - \theta^T x_n)^2}{2\sigma} \right)$$

$$= \sum_{n=1}^{N} \left(\frac{-1}{2} \log (2\pi\sigma^2) - \frac{(y_n - \theta^T x_n)^2}{2\sigma} \right) \quad -(2)$$

Prior is given by

$$P(\theta) = \frac{1}{(2\pi)^{D/2}} e^{-\frac{\lambda}{2}\theta^T\theta}$$

λ is a positive scalar

$$\log P(\theta) = \log \left(\frac{1}{(2\pi)^{D/2}} e^{-\frac{\lambda}{2}\theta^T\theta} \right)$$

$$= \frac{-D}{2} \log (2\pi) - \frac{\lambda}{2} \theta^T\theta \quad -(3)$$

Date
DELTA Pg No.

Substituting values of (2) and (3) in (1) and ignoring the constant values that do not depend on θ we get

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left(\frac{-\lambda}{2} \theta^T \theta - \sum_{n=1}^{N} \frac{(y_n - \theta^T x_n)^2}{2\sigma} \right)$$

Changing the signs to positive and argmax to argmin

$$= \underset{\theta}{\operatorname{argmin}} \left(\frac{\lambda \theta^T \theta}{2} + \sum_{n=1}^{N} \frac{(y_n - \theta^T x_n)^2}{2\sigma} \right)$$

This is the equation for L2 regularization.