ANUSHA MUPPALLA

16286311

TSL HW-2 Assignment.

ull hypothesis means, that there is no combination of individual features (or) parameter like TV, radio, newspaper or internet advertisement will influence sales.

$sales\,c = TV(0) + radio(0) + Newspaper(0) + internet \times 0 + incept(\beta_1)y$

the p value fo each term tests null hypothesis that the coefficient is equal to zero i,e no effect on them.

→ A p value (<0.05) indicates that you can reject (null hypothesis)

⊙ A prediction that has lower p value likely to be a meaningful addition to your model because changes in features valu are related to response value

→ from table p values are significant for (TV) $p<0.001$ and internet ($p<0.0001$) where as it is not significant for radio and newspaper because of large p values so, we can reject TV, internet as they are null hypothesis and no effect what matters

→ But we cannot reject the radio and newspaper as they do affect the advertising as it is a inference problem, when internet and TV are not going up.

5) a.

KNN classification

i) It is a KNN method using k-Nearst neighbor but the final result i,e o/p of y(response) is qualitative i,e majority class of k-Nearst neighbor

ii) knowing neighbours as x_0 then estimating conditional probability $P(y=j/x=x_0)$ for j class

KNN regression

i) It is also using k-NN but the final response is quantitative one for $F(x)$ i.e as average value of corresponding y's in k-nearst neighbour of X.

ii) knowing neighbours then estimating $f(x_0)$ as the average of all training responses.

parametric approach:-

$\hat{y} = \beta_0 + \beta_A x_A + \beta_w x_w + \beta_G x_G + \beta_{Aw} \cdot x_{Aw}$

$\hat{y} = 50 + 0.5 x_A + 1.0 x_w + (-40) x_G + (0.01) x_{Aw} + (1.0) x_{AG}$

where $x_A = 30$, $x_w = 150$, $x_G = \text{male} = 0$

$\hat{y} = 50 + (0.5)(30) + (1.0)(150) - 40(0) + (0.01)(30 \times 150) + (1.0)(30(0))$

$\hat{y} = 260$

b) $\hat{y} = \beta_0 + \beta_A x_A + \beta_w x_w + \beta_G x_G + \beta_{Aw} x_{Aw}$

same as ⓐ question where $x_G = 1$ (female)

$x_A = 30$,

$x_w = 150$

$\hat{y} = 50 + 0.5(30) + (1.0)(150) - 40(1) + (0.01)(30 \times 150) + (1.0)(30)$

$\hat{y} = 250$

c) $x_A = 60$, $x_w = 150$, $x_G = 0$ (male)

$\hat{y} = 50 + 0.5(60) + (1.0)(150) - 40(0) + 0.01(60 \times 150) + (1.0)(60 \times 0)$

$= 50 + 30 + 150 - 0 + 90 + 0$

$\hat{y} = 320$

d) $x_A = 60$, $x_w = 150$, $x_G = 1$ (female)

$\hat{y} = 50 + 0.5(60) + (1.0)(150) - 40(1) + 0.01(60 \times 150) + (1.0)(60 \times 1)$

$\hat{y} = 340$

e) $x_G = 1, x_G = 0 \Rightarrow x_A = ?\ \ x_W = 150$

$\hat{y} = \cancel{50} + (\cancel{0.5})(x_A) + (1.0)(\cancel{150}) - 40(1) + 0.01(x_A * 150) + (1.0)(x_A \times$

$\Rightarrow \cancel{50} + (0.5)(\cancel{x_A}) + (\cancel{1.0})(150) - 40(0) + 0.01[x_A * 150) + (1.0)(x_A$

$x_A - 40 = 0$

$\boxed{x_A = 40}$

Age of 40 for both gender have same blood levels

$y = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$

without knowing more about training data, it is difficult to know which training Rss is lower b/n linear or cubic but if relationship is consider as linear then least squares line to be close to true regression line

smaller than training Rss

b) Answer to the previous question

To this case, test Rss depends upon test data so as of now we have not enough information to decide but we may assume that polynomial regression will have higher test Rss as the overfit and more error than linear regression

c) if relationship is not linear, I would know cubic regression to better fit the training data due to increased flexibility. cubic regression may have some so I can't judge that non linearity

d) It is difficult to judge wheter non linear or cubic regression will lead to a lower Rss in the data (test) as it depends on the true relationship of X and Y. If it is far from linear, cubic regression will lead to lower Rss in test data and better representation of data generating process

Given table have

$n=4$ (no of samples)

$b=5$ (no of groups)

mean of total sum of squares (SS)

$$[T] = \frac{(7+6+8+9+4+3+2+4+11+10+10+10+5+4+5+4+6+5+7+6)}{20}$$

$$= \frac{(30+13+41+18+24)^2}{20} = \frac{(126)^2}{20}$$

$$= \underline{\underline{793.8}}$$

mean b/n groups

$$[B] = \frac{(\Sigma x_1)^2 + (\Sigma x_2)^2 + (\Sigma x_3)^2 + (\Sigma x_4)^2 + (\Sigma x_5)^2}{b-1}$$

$$= \frac{(7+6+8+9)^2 + (4+3+2+4)^2 + (11+10+10+10)^2 + (5+4+5+4)^2 + (6+5+7+6)^2}{4}$$

$$= \frac{3650}{4}$$

$$= \underline{\underline{912.5}}$$

mean within groups.

$$[W] = \frac{7^2+6^2+8^2+9^2+4^2+3^2+2^2+4^2+11^2+10^2+10^2+10^2+5^2+4^2+5^2+4^2+6^2+5^2+7^2+6^2}{15(4-1)}$$

$$= 924$$

Source	Expression	degree of freedom	SS	mean square Error	F	pvalue
B/n groups	[B] = 912.5	5 - 1 = 4	[B] - [T] 118.7	$\frac{SS}{df}$ = 29.675	38.74	F-table
within groups	[W] = 924	20 - 5 = 15	[W] - [B] = 11.5	0.766		
Total	[T] = 793.8	20 - 1 = 19	130.2 (W) - (T)			

pvalue = 0.000000009712

pvalue < 0.0001

It is significant at P<0.1