DATE

Conceptual Questions

① In table 3.4, the p-value of TV, radio and newspaper are < 0.001, < 0.001 and 0.8599 respectively. P-values for TV and radio are significant and hence there is a relationship between TV-Sales and Radio-Sales. However, there is a very large p value for newspaper hence no relationship between Newspaper-Sales.

If the p-value is very less, we say p is significant and we reject null-hypothesis. And null hypothesis is that there is no relationship between a predictor and response. Simultaneously we accept Alternative hypothesis.

② KNN classifier

used for classification. for a given x_0, it selects k points which are closest to x_0 and outputs the label contained by maximum of the k points.

KNN Regression

used for prediction of quantative value, it selects k points closest to x_0 and returns the average of all the values.

DATE

③ (a) Sales = $\hat{\beta}_0 + \hat{\beta}_1 GPA + \hat{\beta}_2 IQ + \hat{\beta}_3 Gender + \hat{\beta}_4 GPA * IQ + \hat{\beta}_5 GPA * Gender$

Sales (Male) = $\hat{\beta}_0 + \hat{\beta}_1 GPA + \hat{\beta}_2 IQ + \hat{\beta}_4 GPA * IQ$

Sales (female) = $\hat{\beta}_0 + \hat{\beta}_1 GPA + \hat{\beta}_2 IQ + \hat{\beta}_3 + \hat{\beta}_4 GPA * IQ + \hat{\beta}_5 GPA$

$= \hat{\beta}_0$

Sales (Female) - Sales (male)

$= \hat{\beta}_3 + \hat{\beta}_5 GPA$

$= 35 - 10\ GPA \longrightarrow (A)$

If the eq (A) is +ve then female earn more on average than male.

Looking at all conditions -

(iii) when GPA is high enough eq (A) is -ve hence males earn more on average than females provided that GPA is high enough.

(b) Sales = $\hat{\beta}_0 + \hat{\beta}_1 GPA + \hat{\beta}_2 IQ + \hat{\beta}_3 Gender + \hat{\beta}_4 GPA * IQ + \hat{\beta}_5 GPA * Gender$

Sales = $50 + 20\ GPA + 0.07\ IQ + 35\ GENDER + 0.01\ GPA * IQ - 10\ GPA * GENDER$

Sales (females) = $50 + 20(4) + 0.07(110) + 35 + 0.01(4)(110) - 10(4)$

$= 7.1 + 50 = 137.1$

PAGE

DATE

(c) false, for significance the value of of coefficients is not the best choice so, we can't conclude.

(4) (a) Complexity $\propto \frac{1}{\text{training error}}$

for training data, RSS decrease with increase in complexity of model. for cubic regression the model will overfit hence for cubic regression RSS will be lower than RSS for linear regression.

(b) for test data, the cubic model will not be able to predict as better as linear regression and RSS for test data ~~will be~~ for linear regression will be lower.

(c) The RSS for cubic regression will be lower. Reason is stated in part (a).

(d) As the model is unknown so it is difficult to define RSS of which model will be lower.
As the nature of model becomes more non-linear test RSS for cubic will be lower.

(5) $\hat{y}_i = x_i \hat{\beta}$ (Given) where,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i y_i)}{\sum_{i'=1}^{n} x_{i'}^2} \quad \because \text{(Given)}$$

To, Prove $\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$

Proof -

$$\hat{y}_i = x_i \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j'=1}^{n} x_{j'}^2}$$

$$= \frac{\sum_{j=1}^{n} x_i x_j y_j}{\sum_{j'=1}^{n} x_{j'}^2} = c_{i'} \sum_{i=1}^{n} x_i y_i$$

$$= \sum_{i=1}^{n} c_{i'} x_i y_i$$

$$= \sum_{i=1}^{n} a_{i'} y_i$$

Hence Proved

where $a_{i'} = \sum_{i=1}^{n} c_{i'} x_i$ where $c_i = \frac{x_i}{\sum_{i'=1}^{n} x_{i'}^2}$

DATE

(6) The least Square line in case of Simple regression is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \longrightarrow (i)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \longrightarrow (ii) \text{ (Given)}$$

putting (ii) in (i)

$$y = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x$$

Now for $x = \bar{x}$

we get.

$$y = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}$$

$$y = \bar{y}$$

Hence $y = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through $(\bar{x}, \bar{y})$

(7) To prove $R^2 = cor(X, Y)^2$ when $\bar{x} = \bar{y} = 0$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= 1 - \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\sum_{i=1}^{n} y_i^2} \quad \text{As } (\bar{y} = 0)$$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 0$$

So putting β_1, β_0 in R^2

DATE

$$= 1 - \frac{\Sigma (y_i - \beta_1 x_i)^2}{\Sigma y_i^2} \qquad \text{(Solving ahead)}$$

$$= \frac{2\frac{(\Sigma x_i y_i)^2}{\Sigma x_i^2} - \frac{(x_i y_i)^2}{\Sigma x_i^2}}{\Sigma y_i^2}$$

$$= \frac{(\Sigma x_i y_i)^2}{\Sigma y_i^2 \Sigma x_i^2}$$

$$(Cor(x,y))^2 = \frac{(\Sigma x_i y_i)^2}{\Sigma x_i^2 \Sigma y_i^2} \qquad \text{(when } \bar{x} = \bar{y} = 0\text{)}$$

Hence $R^2 = (Cor(x,y))^2$