PAGE No.
DATE / /

DADM
Akhil Patil

Q1 Problem 1

(b)

⇒ We have following equations,

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_j y_j^2}$$

With $\hat{y}_i = \hat{\beta}_i x_i$ we may write

$$R^2 = 1 - \frac{\sum_i \left(y_i - \sum_j x_j y_j / \sum_j x_j^2 x_i\right)^2}{\sum_j y_j^2}$$

$$= \sum_j y_j^2 - \left(\sum_i y_i^2 - 2\sum_i y_i \left(\sum_j x_j y_j / \sum_j x_j^2\right) x_i + \sum_i \left(\sum_j x_j y_j / \sum_j x_j^2\right)^2 x_i^2\right)$$

$$\sum_j y_j^2$$

$$\therefore R^2 = \frac{2\left(\sum_i x_i y_i\right)^2 / \sum_j x_j^2 - \left(\sum_i x_i y_i\right)^2 / \sum_j x_j^2}{\sum_j y_j^2}$$

$$= \frac{\left(\sum_i x_i y_i\right)^2}{\sum_j x_j^2 \sum_j y_j^2} = \text{Cor}(x, y)^2.$$

Q4 b) Case Study -

⇒ Hawkins Model's Overview

By Spliting the dataset, we have

Training set :

No Breast cancer = 70/100 * 900 = 630.
Breast cancer = 70/100 * 100 = 70.

Test set:

No Breast cancer = 0.3 x 900 = 270
Breast cancer = 0.3 x 100 = 30.

In order to get what went wrong with the prediction model of Hawkins Let's consider a single case of one women.

Say,

We have an Assumption H, that women is suffering from Breast pain and not Breast cancer but another assumption Ho says women was suffering from Breast cancer.

if Assumption Ho is true (positive) - Breast cancer.

else Assumption Ho is false (negative) - No Breast cancer.

PAGE No.
DATE / /

The table below represents what happens if this other assumption Ho is true or not.

	Null Hypothesis (Ho) valid: Breast cancer	Null Hyp (Ho) is invalid: NO Breast cancer
Accept Ho: Women has Breast cancer.	TP - Women has Breast cancer.	FP - False Alarm, Women has temporary Breast pain but NO Breast cancer.
Reject Ho: Women doesn't have Breast cancer.	FN - Women is healthy but she's dying of Breast cancer.	TN - Women does not have Breast cancer.

Where,
TP - True positive, FP - False Positive
FN - False Negative TN = True Negative.

- Hawkins Model Prediction Results :-

After training the model with 70% of dataset Hawkins scientist tested model with 30% to evaluate the model for its accuracy Their model got 270 predictions right out of 300.

Hawkins accuracy = 270/300 = 0.9

What went wrong then,

By evaluating the model we get,

False Negative = 30
True Positive = 0
True Negative = 270
False Positive = 0

Building a confusion matrix with above data,

	Actual-cancer	Actual-Not cancer	Total
Predicted - cancer	TP=0	FP=0	0
Predicted - Not cancer	FN=30	TN=270	300
Total	30	270	300

In summary,

Hawkins model correctly classified 270 womens who do not have breast cancer as "No Breast Cancer" while it incorrectly classified 30 women who have breast cancer as "No Breast cancer".

Here, the model has convincently classified all the test data as "No Breast Cancer" Accuracy is 90%.

PAGE No.
DATE / /

But,
None of the 'Breast cancer' data is correctly labeled. We can evaluate this by checking Precision & Recall.

By observing the original results we can indicate that people who have Breast cancer the model isn't precise model.

We can prove this by,

Precision = TP / (TP + FP)
Precision = 0 / (0+0) = 0
Precision in % = 0%.

Recall = TP / (TP + FN)
Recall = 0 / (0+30) = 0
Recall in % = 0%.

What this means is that, this model will always classify any data passed into it as 'NO Breast cancer'.

This explains why some patients were shown symptoms of Breast cancer.

Hawkins model did not work.

PAGE No.
DATE / /

Q1 a)

As per table,
The Null hypothesis indicates that advertising budgets of TV, Newspaper or radio do not have any effect on sales.
More precisely,

$H_0^{(1)}: \beta_1 = 0$
$H_0^{(2)}: \beta_2 = 0$
$H_0^{(3)}: \beta_3 = 0$.

The corresponding p-values are highly significant for 'TV' & 'radio' & not significant for 'Newspaper'.
So, we reject $H_0^{(1)}$ & $H_0^{(2)}$ & not $H_0^{(3)}$.
We can conclude that Newspaper advertising budget do not affect sales.

Q4 a)

$\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$

Substituting β values in eq of predicted probabilities;

$$\hat{p}(X) = \frac{e^{-6+0.05x_1+x_2}}{1+(e^{-6+0.05x_1+x_2})}$$

$$= 0.3775$$

Q4.

iii)

eq of predicted probabilities tells us that,

$$= \frac{e^{-6+0.05x_1+x_2}}{(1+e^{-6+0.05x_1+3.5}}$$

$$= 0.5$$

which is equivalent to,

$$e^{-6+0.05x_1+3.5} = 1$$

Taking log on both sides,

$$x_1 = \frac{2.5}{0.05} = 50.$$

Student should study 50 hours.