# DATA MINING ASSIGNMENT 2

Name : Pavani Rangineni

CWID: A20516359

1.1

The line equation is y= $\beta 0$ + $\beta 1$ ,

so substitute $\bar{x}$ for x

y = $\beta 0$ + $\beta 1$ $\bar{x}$ ,

$\beta 0$ = $\bar{y}$ - $\beta 1$ $\bar{x}$

y = $\bar{y}$ - $\beta 1$ $\bar{x}$+ $\beta 1$ $\bar{x}$ = $\bar{y}$

We may conclude that the least square line passes through the point ($\bar{x}$,$\bar{y}$)

1.2

Exercise 3.7 (1)

The null hypotheses associated with table 3.4 that advertising budgets of “TV”, “radio” or “newspaper” do not have effect on sales. The corresponding pvalues are high for “TV” and “radio” and not significant for “newspaper”. Thus we may conclude that newspaper advertising budget do not affect sales

Exercise 3.7 (3) (a)

The least square line is given by y^=50+20GPA+0.07IQ+35Gender+0.01GPA×IQ−10GPA×Gender

y^=50+20GPA+0.07IQ+35Gender+0.01GPA×IQ−10GPA×Gender

which becomes for the males y^=50+20GPA+0.07IQ+0.01GPA×IQ,

y^=50+20GPA+0.07IQ+0.01GPA×IQ,

for the females y^=85+10GPA+0.07IQ+0.01GPA×IQ

.y^=85+10GPA+0.07IQ+0.01GPA×IQ.

So, the starting salary for males is higher than for females on average if 50+20GPA≥85+10GPA50+20GPA≥85+10GPA which is equivalent to GPA≥3.5GPA≥3.5. Therefore 3 is the right answer.

Exercise 3.7 (3)(b)

Predict the salary of a female with IQ of 110 and a GPA of 4.0. we obtain y^=85+40+7.7+4.4=137.1, y^=85+40+7.7+4.4=137.1 which gives us a starting salary of 137100$

Exercise 3.7 (3)(c)

False. To verify the GPA/IQ has an impact on the quality of the model we need to test the hypothesis H0: β4^=0H0: β4^=0 and look at pvalue associated with the $\bar{t}$or the $\bar{F}$ statistic to draw a conclusion.

Exercise 3.7 (4)(a)

It is challenging to determine whether training RSS is lower between linear and cubic without knowing more specifics about the training data. The RSS for the linear regression, however, may be lower than for the cubic regression because the underlying relationship between X and Y is linear, thus we can anticipate the least squares line to be close to the true regression line.