CS 422 HW2 (A20516822)

Monday, 19 September 2022   3:13 PM

Exercise 1.1 :- Section 3.7 - Exercise 6

Using (3.4), argue that in the case of linear regression, the least squares line always passes through $(\bar{x}, \bar{y})$.

Answer :-

Using (3.4), we can say that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\& \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Now, the least squares line is given by

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Lets input x_i as $\bar{x}$

$$\therefore y_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

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

$$\therefore y_i = \bar{y}$$

Since, inputting $\bar{x}$ in the least squares line equation gives us y_i as $\bar{y}$, we can conclude that $(\bar{x}, \bar{y})$ is a point on the least squares line. In other words, least squares line passes through $(\bar{x}, \bar{y})$.

Exercise 1.2 - part I (Section 3.7 - Exercise 1)

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions can you draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of coefficients of the linear model.

Answer :-

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

TABLE 3.4. *For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.*

The sales for this advertising data can be given by

$$\text{Sales }(Y) = \beta_0 + \beta_1 \times TV + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

Now, from table 3.4, we can say that the null hypothesis for TV is that advertising money spent on TV will not have any effect on sales compared to the effect on sales via advertising money spent on radio and newspaper.

The null hypothesis for radio is that advertising money spent on radio will not have any effect on sales compared to the effect on sales via advertising money spent on TV and newspaper.

The null hypothesis for newspaper is that advertising money spent on newspaper will not have any effect on sales compared to the effect on sales via advertising money spent on radio and TV.

The p-values from table 3.4 for TV is 0.0001, radio is 0.0001, and newspaper is 0.8599.

The low p-values for TV and radio tells that both these predictors are statistically significant in predicting the value of sales. Hence, for both of these, we can reject the null hypothesis.

For newspaper, the high p-value tells us that this predictor is not statistically significant in predicting the value of sales. Hence, for newspaper, we will accept the null hypothesis.

Exercise 1.2 - part II (Section 3.7 - Exercise 3)

Suppose we have a data with five predictors,
X_1 = GPA,
X_2 = IQ,
X_3 = Level (1 for college and 0 for high school)
X_4 = Interaction between GPA and IQ
X_5 = Interaction between GPA and Level.

The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

a) which answer is correct, and why?

(i) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.
(ii) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.
(iii) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that GPA is high enough.
(iv) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that GPA is high enough.

Answer :-

The linear regression equation can be given by

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5$$

$$\therefore Y = 50 + (20 \times GPA) + (0.07 \times IQ) + (35 \times Level) + (0.01 \times GPA \times IQ) + (-10 \times GPA \times Level)$$

Considering college graduates, Level = 1,

$$\therefore Y = 50 + (20 \times GPA) + (0.07 \times IQ) + 35 + (0.01 \times GPA \times IQ) - (10 \times GPA)$$

Considering high school graduates, Level = 0,

$$\therefore Y = 50 + (20 \times GPA) + (0.07 \times IQ) + 0 + (0.01 \times GPA \times IQ) - 0$$

$$\therefore Y = 50 + (20 \times GPA) + (0.07 \times IQ) + (0.01 \times GPA \times IQ)$$

We can see that based on $\hat{\beta}_5$, the salary is being confined depending on level of GPA.

To get the same salary as high school graduates fixing IQ, equating both salary values.

$$50 + 20 \times GPA + 0.07 \times IQ + 35 + 0.01 \times GPA \times IQ - 10 \times GPA$$
$$= 50 + 20 \times GPA + 0.07 \times IQ + 0.01 \times GPA \times IQ$$

$$\therefore 35 - 10 \times GPA = 0$$
$$\therefore GPA = 3.5$$

With GPA > 3.5, high school graduates earn more.
Hence, option (iii) is correct which says high school graduates earn more than college graduates provided that GPA is high enough, and for a fixed value of IQ & GPA.

(b) Predict the salary of a college graduate with IQ of 110 and GPA of 4.0

Answer :-

$$Y(\text{salary}) = \hat{\beta}_0 + \hat{\beta}_1 \times GPA + \hat{\beta}_2 \times IQ + \hat{\beta}_3 \times Level + \hat{\beta}_4 \times GPA \times IQ + \hat{\beta}_5 \times GPA \times Level$$

For GPA = 4.0, IQ = 110, Level = 1 (college graduate)

$$Y(\text{salary}) = 50 + 20 \times 4.0 + 0.07 \times 110 + 35 \times 1 + 0.01 \times 4.0 \times 110 + (-10) \times 4.0 \times 1$$
$$= 50 + 80 + 7.7 + 35 + 4.4 - 40$$
$$Y(\text{salary}) = 137.1 \text{ (in thousands of dollars)}$$

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Answer :-

False, we cannot determine the evidence of interaction effect from the coefficient of linear regression. Instead, we need the p-value of this regression coefficient to statistically say whether the interaction effect is significant or not.

Exercise 1.2 - Part III (Section 3.7 - Exercise 4(a))

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Answer :-

Even though the true relationship between X & Y is linear, the cubic regression will give a lower RSS because in cubic regression we are providing another variables to a least square equation which allows to fit the training data tighter, resulting in the poor prediction power of the model when using for test data.