Q(1) (a)

Here is the multiple linear regression of Advertising dataset,

$$Sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \varepsilon$$

Interpretation of CI for β_1 :

→ Here we can see the interval of β_1 provided (for 95% CI) is :

$$[0.0430, 0.0485]$$

The 95% confidence interval says that, that we are 95% confident that the value of β_1 lies between the above interval.

→ So, we can say that for every additional unit of spending on TV advertising, we are 95% confident that the sales increase between 0.043 and 0.048 units, considering all other variables (Radio, Newspaper) to be constant.

→ Therefore, as the overall interval is positive we can conclude that TV advertising has a positive impact on sales.

Conclusion:

Based on the above interpretation we can say that increasing the TV advertising budget will result in increasing sales.

(b) Based on the 95% CI for newspaper coefficient, $[-0.126, 0.0105]$

we have explain why we would fail to reject the null hypothesis $H_0: \beta_3 = 0$ in favor of the alternate hypothesis $H_1: \beta_3 \neq 0$ at the significant level of $\alpha = 0.05$.

We check if 0 is within the CI. In this case, the CI includes 0, which means that a value of 0 for β_3.

→ Since 0 is within the range of coefficient β_3. It means β_3 could be 0, which means there is a possibility that newspaper advertising has no effect on sales.

→ Null Hypothesis (H0): $\beta_3 = 0$

Newspaper advertising has no effect on sales

Alternate Hypothesis (H_1): $\beta_3 \neq 0$.

Newspaper advertising has some effect either positive or negative.

→ In hypothesis testing we reject null hypothesis if the confidence interval doesnot contain 0, which indicates significant relationship.

→ As the CI does include 0, we donot have enough evidence to reject the null Hypothesis.

∴ At $\alpha = 0.05$ (CI = 95%), we fail to reject the null Hypothesis. So, we can say Newspaper advertising has a non-zero effect on sales.

Q(2) (a)

Given regression model

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

→ β_1 is the average or expected change in y for one unit increase in x, instead of just the change in y because of one of the other factors that affect y, error term (ε) which are not related to x.

→ With single observation, the change in y might be different due to other factors by with average observations, the $E(\varepsilon)$ becomes 0, by which we can say that the average change in y, is related to β_1

(b) In multi Linear Regression,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

β_1 is the average change in y, just like in linear regression.

But here we have to consider other variables constant (x_2) while increasing one-unit in x_1.

→ So, we can conclude, β_1 shows effect of x_1 on y by itself without being influenced by other factors (x_2).

Q(3)(a)

→ In first model, when age predictor is present, the R^2 was 0.7501. for y which is muscle mass.

→ for the second regression model, estrogen predictor is added, then the adjusted R^2 increased to 0.8223.

→ So we can conclude that,
As we observe the increase in adjusted R^2, which implies that Estrogen is a meaningful predictor that help us to understand the model better and explain the variation of muscle mass, beyond what is explained by Age alone.

→ Adjusted R^2 penalizes the inclusion of predictors which are not necessary.

(b)

now the third model,

$$\text{muscle mass} = \beta_0 + \beta_1 \text{sleep} + \beta_2 \text{protein} + \beta_3 \text{exercise} + \varepsilon$$

→ this model includes predictors like sleep protein and exercise. which increased the adjusted R^2 to 0.8511 as compared to second model.

→ As we already discussed above, we know that the increase in Adjusted R^2 implies, the predictors help to explain more variations in muscle mass than the second model.

→ We can agree on new predictors providing more variations. which can also say that both new (sleep, protein, exercise) and old (age, estrogen) predictors are important contributors to explaining muscle mass

Q(5)

Yes. if you compare the lengths of the 95% CI (confidence interval) and PI (prediction interval). We can notice that the prediction interval is longer.

Let's get to know about it with simple terms!

what is confidence Interval (CI)?

This is nothing but trying to predict the class average score. we might say "the averge score is in between 75 to 85".

we can observe, the difference in between these figures is less,

So, the CI tells us the certainity about the group performance.

Now lets talk about Prediction Interval.

This is like predicting the score of a specific student. Let's say that

- I am 95% sure that student C score should be around 60 to 100!

→ Here we can see the range is larger or the gap is bigger. As that student can perform better or worse than avg.

Therefore,

Prediction interval is longer as it predict individual outcome which has more uncretainity to predict average outcome.

But confidence interval is smaller as it predicts the average of the outcomes which as less uncretainity.