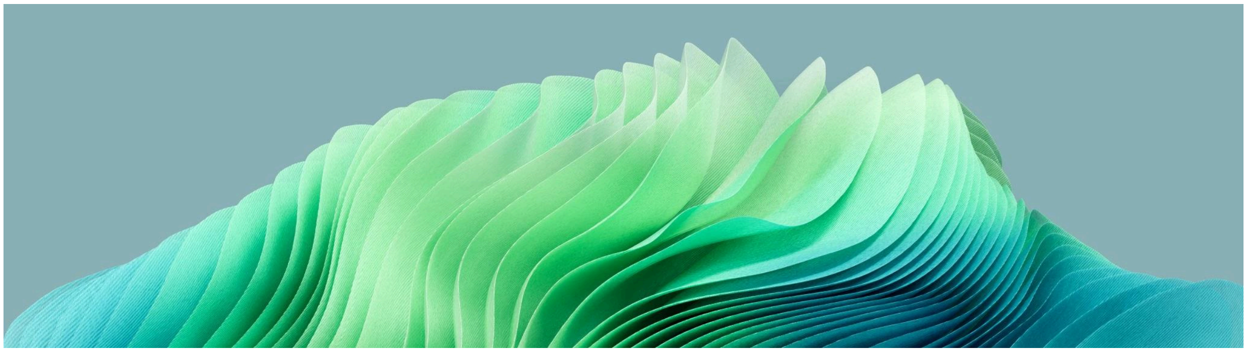
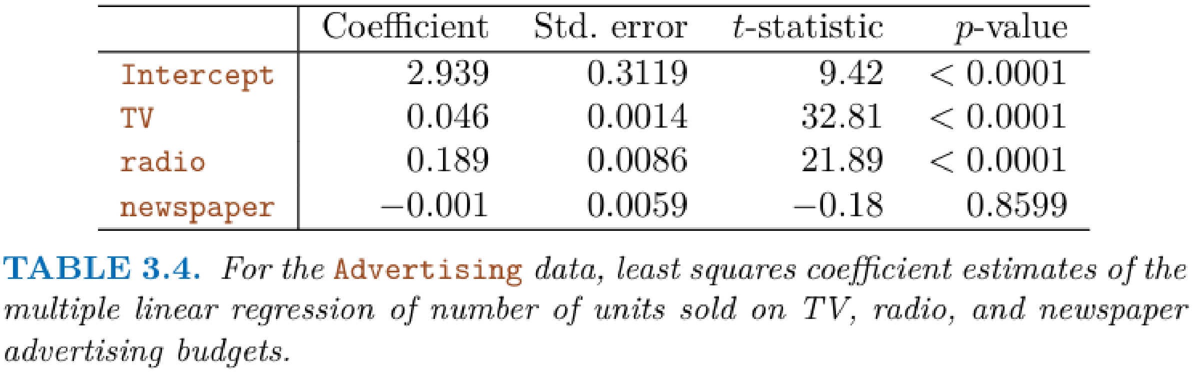


Chapter 3 : Linear Regression

1.Describe the null hypotheses to which the p-values givenin Table3.4 correspond. Explain what conclusions you can draw based on these p-values.Your explanation should be phrased in terms of sales,TV, radio,and newspaper, rather than in terms of the coefficients of the linear model.



	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.*

As we see above in table 3.4, the p values for all the predictors except newspaper is <0.05 making the values obtained by them , (Coefficient, Std.Error , t-stat) statistically significant, Whereas for newspaper , p value>0.05 , meaning there is insufficient evidence to conclude that a relationship between the predictor ‘newspaper’ and the response variable sales exists. A p value dictates that how much of the result one obtained from an experiment was by chance or random error , this difference arises as the sample and population datasets are different , so there can be more than 1 sample for a population and all the different samples can predict different outcomes. Hence having a lesser p value(<0.05) means that our null hypothesis of the predictor and response variable having no relationship among them (beta=0){coefficients in this case} is false and there is in fact a relationship among them which is not by chance.

2. Carefully explain the differences between the KNN classifier and KNN regression methods.

The major difference lies between what kind of **problem is being solved or Our Goal** , if the response variable is qualitative then we use a classifier , if it is quantitative then we use regression. | Both the methods are used for prediction and both are Non Parametric as well
A lower K value corresponds to a Model that has low bias but high variance due to it being highly dependent on just one dataset whereas
A high K value gives a smoother fit with high bias but low variance as now the predicted value is dependent on more than one data point , so even if one changes the other data points stabilizes it. Hence a Propper K value is chosen via Bias Variance tradeOff comparing Test MSE values for regression and error/accuracy test for Classification of different models with different K values

	KNN Classifier	**KNN Regression**
Output Produced	**Qualitative/ Categorical**	**Quantitative / numerical**
How is K used?	The nearest K data points are chosen and a majority rule is applied by default , if the differing variables are the same, then a tie breaker rule is applied.	The average of the closest K datapoints are taken and the predicted value is that average of the nearest K values.

3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = 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 β0=50, ˆβ1=20, ˆβ2=0.07, ˆβ3=35, ˆβ4=0.01, ˆβ5= - 10.

(a)Which answer is correct, and why?

Model formed
Starting sal = 50 + 20 gpa + 0.07 iq + 35 level + 0.01 gpa iq - 10 gpa level
For HS , level =0
Starting Sal = 50 + 20 gpa + 0.07 iq + 0+ 0.01 gpa iq - 0
For College Level = 1
Starting sal = 50 + 20 gpa + 0.07 iq + 35 + 0.01 gpa iq - 10 gpa

CS > HS
CS - HS >0
50 + 20 gpa + 0.07 iq + 35 + 0.01 gpa iq - 10 gpa - 50 - 20 gpa - 0.07 iq - 0.01 gpa iq>0

35-10gpa>0
3.5>gpa
Hence for CS starting sal > HS starting sal their gpa must be < 3.5 , if it is > 3.5 then HS students will get more starting salary

HS>CS
50 + 20 gpa + 0.07 iq + 0+ 0.01 gpa iq - 0 - 50 - 20 gpa - 0.07 iq - 35 - 0.01 gpa iq + 10 gpa >0
-35+10gpa>0
gpa>3.5
Same result as above

i. For a fixed value of IQ and GPA, high school graduates earn more, on average,than college graduates.

True if GPA > 3.5

ii. For a fixed value of IQ and GPA, college graduates earn more,on average, than high school graduates.

True if GPA < 3.5

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

True

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

False

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

For College Level = 1
Starting sal = 50 + 20 gpa + 0.07 iq + 35 + 0.01 gpa iq - 10 gpa
Substituting for given values
Starting sal = 50 + 20 *4 + 0.07 *110 + 35 + 0.01 *4*110- 10 *4
= 50 + 80 +7.7 +35+4.4-40
= 137.1 => $ 137100

(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.

This assumption is false. Even if the coefficient of the interaction term is very small , its magnitude depends on the scale and since it is positive, we know that they have positive synergistic effect. But **whether it needs to be taken into account of not is determined by Hypothesis testing** and comparing **p values** determined by it , if the p value < 0.05 - that means there is an interaction effect and we need to take this effect into account, whereas if p value> 0.05 that means that the null hypothesis has enough evidence to be true and we will not take this effect into account.
Another note to take into account is the **hierarchical principle** which dictates that whether the individual terms like GPA and IQ themselves have high p-value and interaction term has low p value , the interaction term will still be taken into account meaning , IF the interaction term has low p value and is statistically significant , The individual predictors having high p value dont mean anything and we must include them in our model as they are related to the interaction effect which in turn is related to the response variable , therefore leaving them out can alter the meaning of the model.