**Homework 2**

**Question 1.1:**

**Problem 1:**

Answer: The null hypothesis for TV feature is that when both radio and newspaper are present TV feature has no effect on sales attribute. The null hypothesis for Radio feature is that when both tv and newspaper are present Radio feature has no effect on sales attribute. The null hypothesis for Newspaper feature is that when both radio and tv are present Newspaper feature has no effect on sales attribute. From the P values given in the table 3.4 we can infer that P value for newspaper is high (close to 1) so Null hypothesis for Newspaper Feature hold. On the other hand, TV and Radio have low P value which states that Null Hypothesis for Tv and Radio doesn’t hold, and these variables have high correlation towards the sales.

Problem 3:

**a)**

Answer: iii)

Result(y) = 50 + 20(GPA) + 0.07(IQ) + 35(Gender) + 0.01(GPA -> IQ) – 10(GPA -> Gender).

Here Beta3 and Beta5 both have gender. Beta3 is positive and Beta5 is negative. Beta3 won’t have an effect on males as male gender is represented as 0. So once Beta5 multiplier increases after a certain value it will have negative impact for females. 35/10 = 3.5. So, when GPA is greater than 3.5 Male students earn more on average than the female students after graduation.

**b)**

Answer:

Result(y) = 50+20(GPA)+0.07(IQ)+35(Gender)+0.01(GPA ->IQ) – 10(GPA -> Gender).

= 50 + 20(4) + 0.07(110) + 35(1) + 0.01(4\*110) – 10(4\*1)

= 137.1

Result = 137.1.

**c)**

Answer: False

Even though the value of IQ is high compared to the other predictors, the regression coefficients could be small. So, we can’t comment on the interaction effect.

**Problem 4:**

**a)**

Cubic Regression has more number of predictors and it will be a better fit to the data than the linear regression model. So, the Cubic regression model will have low RSS compared to the linear regression model.

**b)**

Since we are using cubic regression for training the model, this might result in overfitting the data during training and having higher RSS during testing. So, the error will be higher for cubic regression model than linear model during testing.

**C)**

The cubic regression model tries to fit all the data as it highly flexible. So, the cubic regression model will have lower RSS than the linear regression model.

**D)**

Since the true model is non-linear and we don’t know how far it is from linearity, we can’t comment on which model produces lower RSS. If the True model as close to linear then linear model will produce lower RSS during testing and vice versa.

**Question 1.2:**

**Problem 4:**

**a)**

Answer: On average we use 10% of the observation to predict the response. If X=0.6 we check in the range [0.55,0.65].

**b)**

Answer: (0.1) ^2 \* 100 = 0.01 \* 100 = 1% on average we use 1% to make the prediction.

**C)**

Answer: (0.1) ^100 \* 100 = 100 / 10 ^100 = 10 ^ -98 % to make the prediction.

**D)**

When the number of features is large the dimensions in representing the data, points is complex, and the data points stay far apart from each other So KNN preforms poorly when the number of dimensions is large.

**E)**

Answer: P = 1, L = (0.1) ^1

P = 2, L = (0.1) ^ (1/2), ………………..., P = N, L = (0.1) ^ (1/N).

As the number of dimensions (p) increases length of the hypercube also increases.

**Problem 6:**

**a)**

P(X) = exp (B0 + B1 X1 + B2 X2) / 1 + exp (B0 + B1 X1 + B2 X2).

= exp (-6 + 0.05 \* 40 + 3.5) / 1 + exp (-6 + 0.05 \* 40 + 3.5).

= 37.75%

**b)**

P(X) = exp (B0 + B1 X1 + B2 X2) / 1 + exp (B0 + B1 X1 + B2 X2).

0.5 = exp (-6 + 0.05 \*X1 + 3.5) / 1 + exp (-6 + 0.05 \* X1 + 3.5).

X1 = 50 hours.