**IML Assignment - 1 Solutions**

**Question 1 Solution**

**(a)**

True- 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.

Let,

* x1 = GPA
* x2 = IQ
* x3 = Level (College - 1, High School - 0)
* x4 = Interaction b/w GPA and IQ (x1.x2)
* x5 = Interaction b/w GPA and Level (x1.x3)
* Salary = b0 + b1x1 + b2x2 + b3x3 + b4x4 + b5x5 = 50 + 20x1 + 0.07x2 + 35x3 + 0.01x4 - 10x5

for fixed IQ and GPA at x1’ and x2’: -

Salary (high school) = 50 + 20x1’ + 0.07x2’ + 35\*(0) + 0.01(x1’.x2’) -10(x1’.0)

= 50 + 20x1’ + 0.07x2’ + 0.01(x1’.x2’)

Salary (college) = 50 + 20x1’ + 0.07x2’ + 35\*(1) + 0.01(x1’.x2’) -10(x1’.1)

= 50 + 20x1’ + 0.07x2’ + 35 + 0.01(x1’.x2’) - 10(x1’)

= Salary (high school) + 35 - 10(x1’)

From here:

Salary (college) - Salary (high school) = 35 - 10x1’

Assuming the salary difference to be more than equal to zero, we get:

35 - 10x1’ >= 0 x1’ <= 3.5

Assuming the salary difference to be less than equal to zero, we get:

35 - 10x1’ <= 0 x1’ >= 3.5

Hence, for a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is more than equal to 3.5

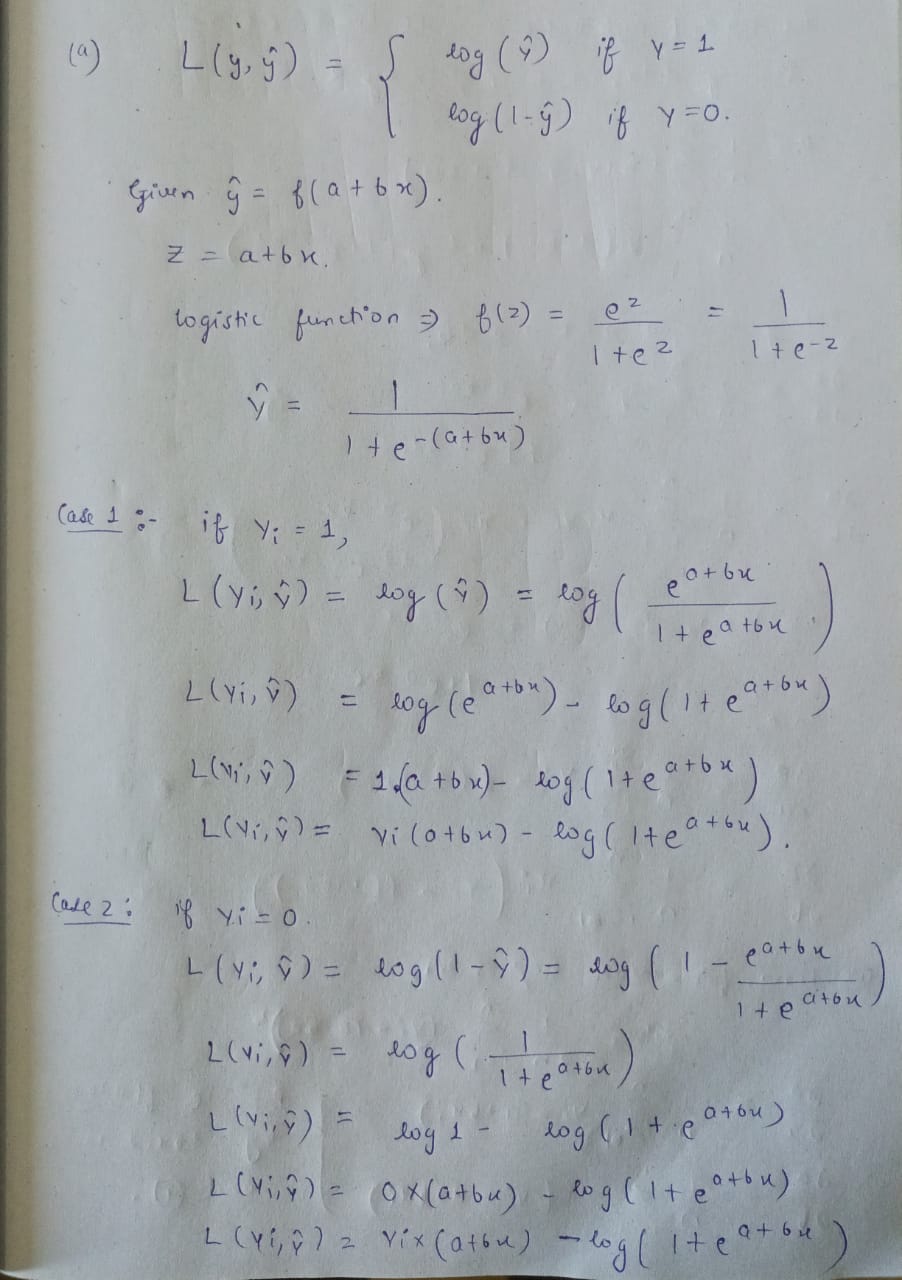
**(b)**

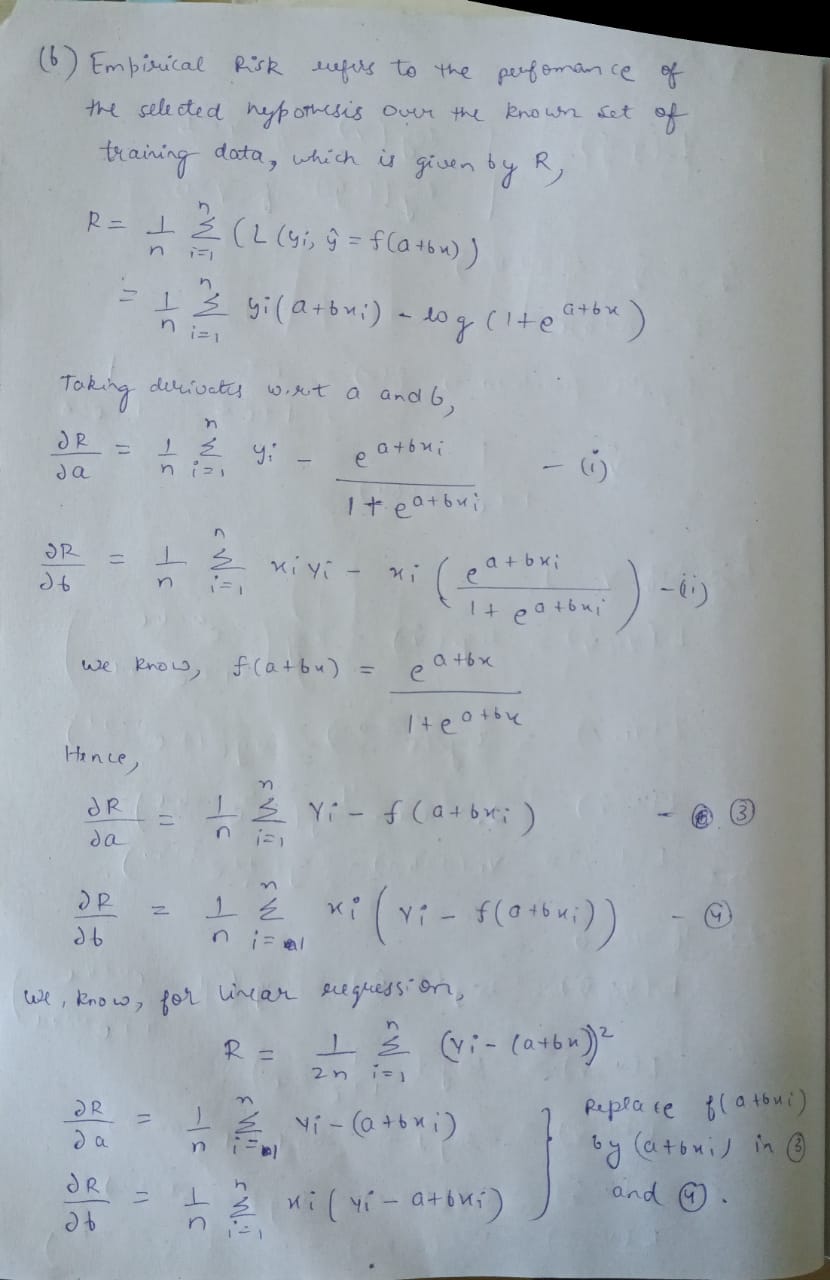
Ans: Salary = 50+20(4)+0.07(110)+35+0.01(110x4)-10(4) = 137.1 (in thousands of dollars)

**(c)**

Ans: False because the magnitude of coefficient is not an indicator of statistical significance

**Question 2 Solution :**

****

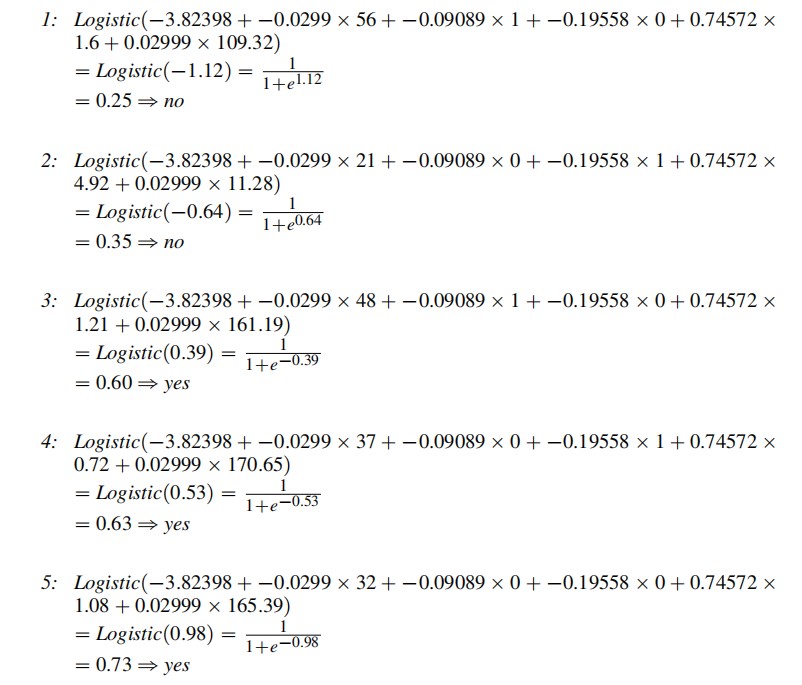
****

**Question 3 Solution :**

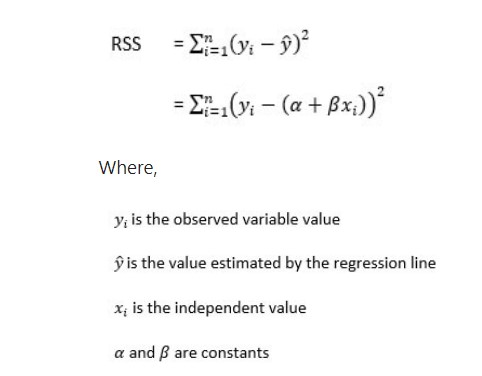
Calculating the predictions made by the model simply involves inserting the descriptive features from each query instance into the prediction model. The only extra thing that must be considered in this case is the categorical descriptive feature SOCIO ECONOMIC BAND.

We can note from the regression equation that this one feature has been expanded into two: SOCIO ECONOMIC BAND B and SOCIO ECONOMIC BAND C. These are binary features, indicating that the original feature was set to the level *b* or *c*. It is assumed that when both of these features are set to 0, then the original feature was set to *a* (the choice of which level to leave out is arbitrary). The other pieces of information required are that the *yes* level is the positive level, and the classification threshold is 0.5.

With this information, the predictions can be made as follows:



**Question 4 Solution :**

****

A lower RSS signifies that the regression model explains the data better, indicating the least variance and better fit.

1. We would expect the training RSS for the cubic model to be lower because it is more flexible (because of the x^2 and x^3 terms), which allows it to fit more closely to the training data. This would reduce the RSS despite it not being representative of a closer approximation to the true linear relationship , that is f(x) = B0 + B1 (x)
2. We would expect the test RSS for the linear regression to be lower, because the assumption of high bias is correct, so lack of flexibility is of no cost in estimating the true f(x). The cubic model, being more flexible is likely to over-fit and is not a true representative of f(x), hence will result in higher RSS.

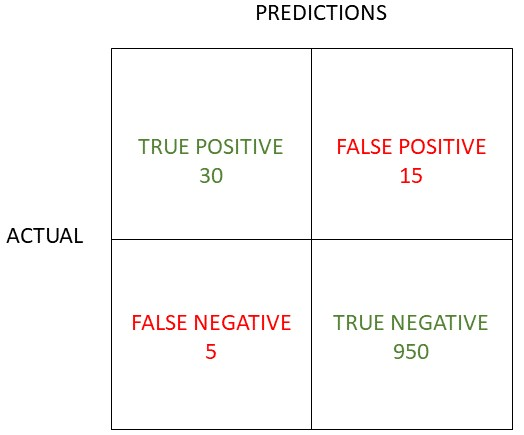
**Question 5 Solution :**

True Positive (TP) a correct positive test – 30

True Negative (TN) a correct negative test – 950

False Positive (FP) an incorrect positive test – 15

False Negative (FN) an incorrect negative test – 5



Accuracy = number of correct predictions / total number of predictions

30+950  /  30 + 15+ 950 + 5 = 980/1000 = 49/50 or 98%

Precision = true positive / true positive + false positive

30 / 30+15 =30/45 =2/3 or 66.7%

Recall (and sensitivity) = true positive / true positive + false negative

30 / 30+5=30/35 = 0.857 or 85.7%

 F1 score = 2 x (precision\*recall  / precision + recall)

= 2 \* (0.57/1.52) = 2\*0.375 =0.75 or 75%