Quizzes : Module 3 Part 2

Linear Regression

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

If two variables have a high positive correlation (above 0.7, say), what can we say about the variables? Multiple options may be correct.

Answer Choice	Verdict	Explanation
There is a strong linear relationship between the variables.	Correct	Right. Strong positive correlation does suggest a strong linear relationship. Causality may not be clear though.
Knowing one of the variables may help us predict the values of the other.	Correct	Strong positive correlation suggests linear relationship, and such a relationship is useful in prediction.
There is no linear relationship between the variables.	Wrong	Nope. Strong positive correlation does suggest a strong linear relationship. There may not be causality though.
There is no non-linear relationship between the variables.	Wrong	Sorry. Can't infer anything about non-linear relations from correlation. It only tells us about linear relation.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 7 and Slide 8

General Comment : Correlation does tell us about linear relationship between variables, but not about causality. Correlation, in the classical sense, does not tell us anything about non-linear relationship between variables. High correlation means strong linear relationship, which is quite useful if we want to predict one variable using the other. This is something that we see frequently in case of Linear Regression. Also note that we are talking about Pearson Correlation Coefficient so far; you may want to check it out on Wikipedia or any book on Statistics. ;-)

Question 2

Suppose that you have 1000 observations in a dataset, and you plan to use 750 as your Train set. Which of the following is/are most appropriate in such a scenario?

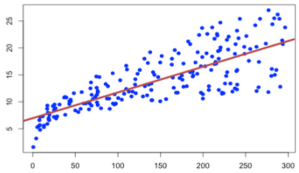
Answer Choice	Verdict	Explanation
The remaining 250 observations (or a subset of these) may be used as the Test set.	Correct	Correct. The remaining 250 observations have not been used for Training, and hence may be used for Test.
The 750 observations in Train set should be selected carefully -- may be chosen uniformly at random from the Dataset.	Correct	Correct. The Train set should be representative of the main Dataset. Uniform random selection (often) helps in achieving that.
It does not matter how we choose the 750 observations from the Dataset to get Train set. Choose the first 750 observations.	Wrong	Actually, it does matter. The Train set should be a representative of the Dataset. You do not know if the "first 750" data points are randomly distributed in the dataset or arranged in some predetermined fashion. Try selecting uniformly at random, it often helps reducing such bias.
It does not matter which observations from the main Dataset are taken for the Test set. Choose any 250 at random.	Wrong	Actually, it does matter. We should not use the same observations for Train and Test. So, the remaining 250 observations are the ones from which we should draw the Test set.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 5

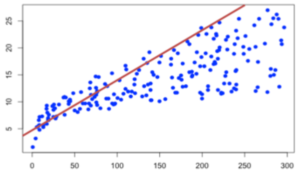
General Comment : There are two crucial aspects in choosing the Train and Test sets. First, the Train set should be a representative of the main Dataset, and hence, should be chosen quite carefully. Uniform random selection from the main labeled dataset often helps achieve this. Second, the Test set and Train set should not have any overlap, as in that case, observations used for Train will again be used for Test, reducing the fairness of the whole learning process.

Question 3

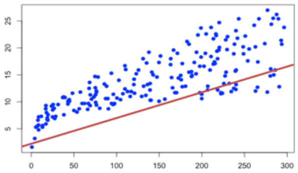
Arrange the following linear models in decreasing order of how well they fit the dataset -- the Best Fit one to the Worst Fit one.



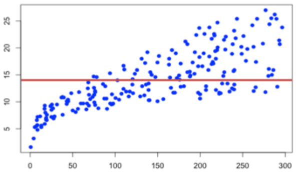
Minimum RSS



Slightly large RSS



Much higher RSS



Maximum RSS (= TSS)

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 10 and Slide 11

General Comment : The best fit linear model will make the least Sum Square of Errors (RSS), while the worst fit linear model will make the most Sum Square Error (RSS). Minimum RSS is close to zero, while worst RSS is equal to TSS.

Question 4

Arrange the steps of Linear Regression, as follows, in order in which they are executed.

1. Guess the initial values of the "Parameters" for the hypothesized Linear Model.
2. Predict the values of the Response Variable for all observations in Train data.
3. Compute the Errors in Train data, compared to actual values of the Response.
4. Choose a specific Cost Function (like Sum Square of Errors) for Optimization.
5. Reassign or Tune the "Parameters" of the model to Optimize the Cost Function.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 11

Question 5

Which ones of the following are decent choices for Cost Function in Linear Regression? Multiple options may be correct.

Answer Choice	Verdict	Explanation
Sum Square of Errors / Residual Sum of Squares over the Train Set	Correct	Of course. This is the most commonly used Cost Function in Linear regression.
Absolute Sum of Errors / Residual Absolute Sum over the Train Set	Correct	Sure, this works too. Note that optimization is a little hard with the non-differentiable function. But still, it will work in theory.
Minimum Absolute Error / Minimum Absolute Residual in the Train Set	Wrong	No, this won't work. Think about it -- even a horribly fit line can pass through one of the points in the Train set. In that case, the minimum error will actually be Zero (0), resulting in the minimum possible Absolute Error. You need to consider all points.
Maximum Squared Error / Maximum Residual Square in the Train Set	Wrong	No, this won't work. Think about it -- all the weight for this Cost Function is placed on the point that is farthest from the line. Thus, the cost function is unnecessarily biased to the outliers, and not to the other points. You need to consider all points.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 11

General Comment : Think about the Cost Function as a cumulative measure of all Errors in the Train set. We also require the Linear Regression to work by "optimizing" the Cost Function, which often requires differentiating the cost function (remember calculus in terms of maximum or minimum of a function?). Put these two ideas together, and think again -- which one of the options are potential Cost Functions.

Question 6

The hypothesized linear model and the cost function in a linear regression problem are as follows. Which of the following are True? Multiple options can be correct.

Linear Model : *Response* = *a* x *Predictor* + *b*

Cost Function : *J* = Sum of (*Response* – *a* x *Predictor* - *b*)^2

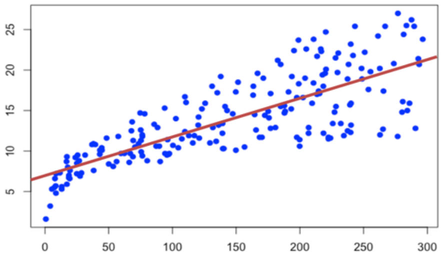
Answer Choice	Verdict	Explanation
a and *b* are the Parameters of the hypothesized Linear Model	Correct	Correct. This is by definition of the Linear Model. To estimate these parameters is the goal of Linear Regression.
Response and *Predictor* in the Cost Function *J* are available as fixed values from the Train set	Correct	Correct. Even though *Response* and *Predictor* look like variables, they are not. All values of these two items are available from the Train set, and we simply plug these values in the Cost Function before the optimization process starts.
a and *b* are the actual Variables in the Cost Function *J* after we put in the values from the Train set	Correct	Correct. When we say that optimizing the Cost Function is the goal of Linear Regression, we mean that the Cost Function is a function of the parameters *a* and *b*, and that we have to optimize (minimize) the bi-variate Cost Function *J*(*a*,*b*).
Cost Function *J* is a function of *Response* and *Predictor*, as the parameters *a* and *b* are guessed.	Wrong	The cost function is really a function of the parameters *a* and *b*. Even though *Response* and *Predictor* look like variables, they are not. All values of these two items are available from the Train set, and we simply plug these values in the Cost Function before the optimization process starts. When we start the optimization, we guess the values of *a* and *b* just to initiate the process.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 11

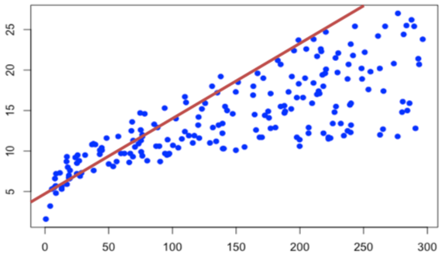
General Comment : Make sure you understand that the variables in a Cost Function are actually the "parameters" of the model, and that the Response and Predictors are all constant values, available from the Train set.

Question 7

The following linear models (A and B) were fit on the Train data shown in the figures. Which one would best predict the Test data?



Model A



Model B

Answer Choice	Verdict	Explanation
It is impossible to say which model will better predict on the Test data, as it depends on other factors.	Correct	Correct. If the Train data is similar to the Test data, Model A will do a good job of prediction, but there is no guarantee otherwise. In Machine Learning, we try hard to make sure that the Train data is similar to the Test data. However, if we do not know for sure that the Train data is similar to the Test data, we can't really say that the model fitting the Train data best will surely be the best predictor for the Test data. It really depends on a lot of other factors.
Linear Model A will best predict the Test data, as it best fits the Train data, as shown in the figure.	Wrong	This would be correct only if the Train data is similar to the Test data. In Machine Learning, we try hard to make that happen. However, if we do not know for sure that the Train data is similar to the Test data, we can't really say that the model fitting the Train data best will surely be the best predictor for the Test data.
Linear Model B will best predict the Test data, as it best fits the Train data, as shown in the figure.	Wrong	It is clear from the figures that Linear Model A best fits the Train data, if we consider standard Sum Squared Errors. Thus, the statement is not correct. However, we do not know for sure if Model B is a better predictor for Test data, as we do not know the Test data yet.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 13 and Slide 14

General Comment : Note that the Train data and Test data are mutually independent, and thus, it is hard to conclude something about the Test data unless we know that the Train data is really similar to the Test data. This is a big issue in Machine Learning, known as "Generalization of Model". Go ahead and Google for this issue. :-)

Question 8

Suppose that there are 100 observations in a Train set, and the Residual Sum of Squares (RSS) for a linear model is 745. What is the Mean Squared Error (MSE) of the linear model on the Train set?

Answer Choice	Verdict	Explanation
7.45	Correct	Correct. The MSE is RSS divided by the number of data points in the Train set, that is 745/100 = 7.45.
74500	Wrong	Nope. It is the other way round -- that is, MSE = RSS divided by the number of data points. Check again.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 14

General Comment : RSS is the Sum of Square of Errors, while MSE is the Mean of Square of Errors. Hence the relation.

Question 9

Suppose the MSE of a linear model on the Train set is 4.25, the number of observations in the Train set is 100, and the Variance of the Response variable in the Train set is 8.5. What is the value of R^2 on the Train set in this case?

Answer Choice	Verdict	Explanation
0.5	Correct	Correct. R^2 = 1 – (MSE/VAR) in the Test set. The information about 100 observations is redundant in this case.
-49	Wrong	Nope. R^2 = 1 – (MSE/VAR) in the Test set. The information about 100 observations is redundant in this case.
-0.5	Wrong	Can’t be negative. Remember that R^2 can only be between 0 to 1.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 14

General Comment : Check Slide 14 (last but one) in this lesson to find out the relationship between R^2, MSE, VAR, RSS and TSS. Remember that R^2 can only be between 0 to 1. Hence, be careful and cross-check your calculation.

Extra note : Computing R^2 on Test Data is a little misleading. You compute R^2 on Train Data to judge the performance of your model, but evaluate the same performance on Test Data using MSE or RMSE, and not using R^2. If you obtained negative R^2 in Test Data during your Lab Exercises, that is possible, as VAR is computed on Train Data.

Question 10

What do you think happens if for some linear model, we get R^2 = 1 in the Train set? Is it good or bad for prediction?

Answer Choice	Verdict	Explanation
R^2 = 1 means the Residual Sum of Squares (RSS) is Zero, that is, the model "perfectly" fits Train set. However, there is no guarantee that it will best predict the Test set, as we do not know if it is similar to the Train set.	Correct	Correct. R^2 = 1 or RSS = 0 on the Train set just means that the model "perfectly" fits the Train set. There is no guarantee that it will best predict the Test set, as we do not know if it is similar to the Train set. In fact, we get a little worried if R^2 = 1 in practice -- it often means that we've "overfit" the Train set. Google "overfitting". :-)
R^2 = 1 means the Residual Sum of Squares (RSS) is Zero, that is, the model "perfectly" fits Train set. This is the ideal case, as we have the best fit model. Definitely, this model will be the best one to predict on the Test set.	Wrong	Well, you are correct on the first count. RSS is really 0 on the Train set. However, this just means that the model "perfectly" fits the Train set. There is no guarantee that it will best predict the Test set, as we do not know if it is similar to the Train set. In fact, we get a little worried if R^2 = 1 in practice -- it often means that we've "overfit" the Train set. Google "overfitting". :-)
R^2 = 1 means the Residual Sum of Squares (RSS) is Maximum (equal to TSS), that is, the model is the "worst" fit on the Train set. This is the worst case, and definitely, the model will be the worst one to predict on the Test set.	Wrong	Wrong. Check the formula for R^2 once again. R^2 = 1 means RSS = 0, not maximum.

Reference	Module 3 Topic 2 : Uni-Variate Linear Regression	No specific slide. It's slightly beyond.

General Comment : Check the formula for R^2 to obtain its relationship with RSS. This should be really easy. However, the second part is non-intuitive. Even if a model "perfectly" fits the Train set, there is no guarantee that it will best predict the Test set, as there is a chance of "overfitting". Go ahead and Google for this issue. ;-)