# WORKSHEET

MACHINE LEARNING – WORKSHEET 11 (LINEAR REGRESSION)

## In Q1 to Q8, only one option is correct, Choose the correct option:

1. What happens to R2 measure if we add a new feature?
   1. remains same B) always increases

C) may or may not increase D) always decreases

**Ans. B) always increases**

1. The correct relationship between SST, SSR and SSE is given by:
   1. SSR = SST + SSE B) SST = SSR + SSE

C) SSE = SSR – SST D) None of the above

**Ans. B) SST = SSR + SSE**

1. Residuals in regression analysis can be defined as:
   1. difference between the actual value and the predicted value.
   2. difference between the actual value and the mean of predicted value.
   3. difference between the actual value and mean of dependent variable.
   4. None of the above.

**Ans. A) difference between the actual value and the predicted value.**

1. In a simple linear regression model, if we change the input variable by 1 unit, then how much output variable will change?
   1. By 1 B) No Change

C) By its slope D) None of the above

**Ans. C) By its slope**

1. If the coefficient of determination is equal to 1, then correlation coefficient:
   1. must also be equal to 1 B) can be either -1 or 1

C) can be any value between -1 to 1 D) must be -1

**Ans. A) must also be equal to 1**

1. Which of the following plot is best suited for the linear relationship of continuous variables?
   1. Scatter plot B) Histograms

C) Pie charts D) All of the above

**Ans. A) Scatter plot**

1. The ratio of MSR/MSE produces:
   1. t-statistics B) f-statistics

C) z-statistics D) None of the above.

**Ans. B) f-statistics**

1. Which of the following regularizations uses only L2 normalization for its penalty parameter?
   1. Lasso B) Elastic Nets

C) Ridge D) All of the above

**Ans. C) Ridge**

## In Q9 to Q11, more than one options are correct, Choose all the correct options:

1. Which of the following statement/s are true for best fitted line?
   1. It shows the causal relationship between dependent and independent variables
   2. It shows the positive or negative relation between dependent and independent variables
   3. It always goes through origin
   4. It is a straight line that is the best approximation of the given data sets

**Ans. B) It shows the positive or negative relation between dependent and independent variables**

**D) It is a straight line that is the best approximation of the given data sets**

1. Regularizations helps in:
   1. Reducing the training time B) Generalizing the test set

C) Automatic feature selection D) Grouping the data

**Ans. A) Reducing the training time, B) Generalizing the test set**

1. Linear regression can be implemented through:
   1. Normal Equation B) Singular Value Decomposition

C) Parity checks D) nodes

**Ans. A) Normal Equation, B) Singular Value Decomposition**

## Q12 to Q15 are subjective answer type questions, Answer them briefly.

1. Explain R2 and adjusted R2 metrics?

**Ans.** R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model.

1. Explain the cost function of linear regression?

**Ans.** It is a function that **measures the performance of a Machine Learning model** for given data. Cost Function quantifies the error between predicted values and expected values and**presents it in the form of a single real number**. Depending on the problem Cost Function can be formed in many different ways. The purpose of Cost Function is to be either:

* **Minimized**- then returned value is usually called **cost**, **loss** or **error**. The goal is to find the values of model parameters for which Cost Function return as small number as possible.
* **Maximized**- then the value it yields is named a **reward**. The goal is to find values of model parameters for which returned number is as large as possible.

1. Differentiate SSE, SSR and SST.

Ans. **SST** - The **sum of squares total**, denoted **SST**, is the squared differences between the observed dependent variable and its **mean**. You can think of this as the dispersion of the observed variables around the [**mean**](https://365datascience.com/measures-central-tendency/) – much like the **variance** in descriptive statistics.

It is a measure of the total variability of the dataset.

**Note**: There is another notation for the **SST**. It is **TSS** or **total sum of squares**.xcdsewfrrrrrrrrrrrrrr

**SSR**-The second term is the **sum of squares due to regression**, or **SSR**. It is the sum of the differences between the predicted value and the **mean** of the dependent variable. Think of it as a measure that describes how well our line fits the data.

If this value of **SSR** is equal to the **sum of squares total**, it means our **regression** **model** captures all the observed variability and is perfect. Once again, we have to mention that another common notation is **ESS** or **explained sum of squares**.

**SSE**- The last term is the **sum of squares error**, or **SSE**. The error is the difference between the observed value and the predicted value.

We usually want to minimize the error. The smaller the error, the better the estimation power of the **regression**. Finally, I should add that it is also known as **RSS** or **residual sum of squares**. Residual as in: remaining or unexplained.

1. What are the various evaluation metrics for linear regression?

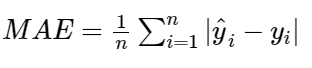
Ans. Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship. Let us look at**MSE, MAE, R-squared, Adjusted R-squared, and RMSE.**

**Mean Squared Error (MSE)**

Image for postThe most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.

MSE penalizes large errors.

**Mean Absolute Error (MAE)**

This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent.

MAE does not penalize large errors.

**R-squared or Coefficient of Determination**

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.

## 

## Root Mean Squared Error (RMSE)

This is the square root of the average of the squared difference of the predicted and actual value.

R-squared error is better than RMSE. This is because R-squared is a relative measure while RMSE is an absolute measure of fit (highly dependent on the variables — not a normalized measure).

**Adjusted R-squared — selection criterion**

The main difference between **adjusted R-squared**and R-square is that **R-squared** describes the amount of variance of the dependent variable represented by every single independent variable, while **adjusted R-squared** measures variation explained by only the independent variables that actually affect the dependent variable.

