**Module 7**

* [Video Transcripts](https://student.emeritus.org/courses/4765/files/3087195?wrap=1)
* [Download Video Transcripts](https://student.emeritus.org/courses/4765/files/3087195/download?download_frd=1)
* [Quick Reference Guide](https://student.emeritus.org/courses/4765/files/3087196?wrap=1)

**Savio’s Session**

Cross-validation:

Y = mx + c -> y dependent, x is independent variable (feature)

sks\_sfs.fit()

boston.feature\_names[sks\_sfs.get\_support()]

**Notes:**

The linear regression equation can be expressed as

 X is the explanatory variable - waiting | total bill

 Y is the dependent variable - duration | tip

exploratory data analysis (EDA)

The categorical data type has two subcategories: nominal and ordinal.

Nominal data (sometimes called ‘labeled’ or ‘named’ data) type is used to name variables with no numerical value and that have no inherent ordering. For example, sex (male or female) has no specific order. This type of data is typically collected via surveys or questionnaires.

Conversely, ordinal data has a clear ordering or scale to it. For example, things such as bug scale, customer satisfaction survey data, or interval scale have different collection and analysis techniques but are all considered ordinal data.

**Absolute Loss**

A synonym for L1 or mean absolute error

**Categorical Data**

A data type that can be divided into two subtypes: nominal or ordinal data

**Huber Loss**

A composite of both MSE and MAE; also known as smooth mean absolute error

**L1 Loss**

A synonym for absolute loss and mean absolute error (MAE)

**L2 Loss**

A synonym for mean squared error (MSE)

**Linear Regression**

A model of the relationship between two variables, fit to a linear equation

**Mean Absolute Error (MAE)**

The average sum of the absolute difference between the actual and the predicted values; also known as L1 loss or absolute loss

**Mean Squared Error (MSE)**

The average squared difference between the predicted and actual variables; also known as L2 loss

**Nominal Data**

Data with no numerical value or inherent ordering; sometimes called “labeled” or “named” data

**Non-Linear Regression**

A model of the relationship between two variables, fit to a non-linear equation

**Ordinal Data**

Non-numeric data that has a clear ordering or scale, such as “Good”, “Better”, “Best”

**Mean squared error (MSE)**

The mean squared error is probably the most popular loss function for regression. It is interpreted using the average squared difference between the predicted and actual variables.

**Advantages**

The MSE algorithm can converge to the minimum efficiently with small errors as the gradient slowly declines.

**Drawbacks**

Squared values generally result in a higher training rate; however, a very large loss may result in a dramatic jump during backpropagation, which is not desirable. MSE is also sensitive to data outliers and, as a result, data outliers may impact your network more heavily, given the significantly higher loss associated with them.

**Mean absolute error (MAE)**

The mean absolute error is the simplest method for calculating loss and can be found by taking the average sum of the absolute difference between the actual and the predicted values.

**Advantages**

MAE is an effective method for calculating loss because it is simple and computationally inexpensive.

**Drawbacks**

MAE calculates loss by taking into consideration all errors on the same scale, meaning all errors are weighted equally while calculating the mean.

**Inference**

A statement about the relationship between the response and the predictor variables

**Overfitting**

A modeling error where a function is too closely fit to a limited set of data points resulting in an inability of the model to generalize to other datasets

**Pipeline**

A scikit-learn functionality that assembles several steps that can be cross-validated together while setting different parameters

**Prediction**

A conclusive statement that pertains to a future event or occurrence

**Test Set**

A subset of the initial dataset that is set aside to test the trained model

**Training Set**

The dataset you use to train an algorithm or machine learning model

**Validation Set**

A subset of the initial dataset that is used to give an estimate of model skill while tuning the model’s hyperparameters

**Module Issues:**

In video 7.1 10kg is supposed to be 22lb, not 2.2lb!

Codio Activity 7.2 link launches 7.1 - Fixed

Codio Activity 7.3 link launches 7.2 - Fixed

Codio Activity 7.3 Problem 2 asking to return an array when a single float is needed.

Codio Activity 7.6 Problem 4, we need to set index to Features: error\_df.set\_index('Features')

**Quizes:**

Regression problems can be solved using linear models. : True

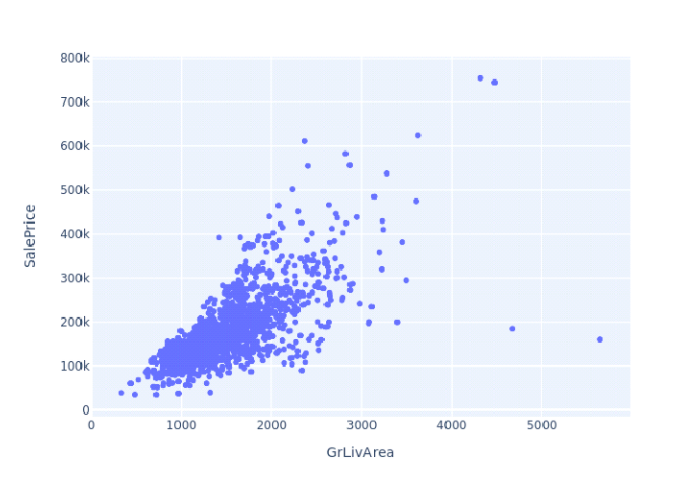
*You are correct! The answer “*True*” is correct because linear models can be used to solve regression problems.*

Given the weight and height of a person, predicting biological sex is a regression problem. : False

*You are correct! The answer “*False*” is correct because the predictor biological sex is not a real valued outcome.*

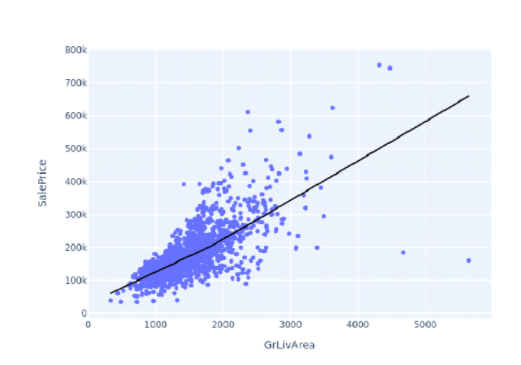
What kind of correlation is there between the two variables shown in the scatterplot? : Positive correlation

*You are correct! The answer “*Positive correlation*” is correct because the graph shows that with the increase in value for one variable the value for the other also increases.*



The trend line in the given scatterplot is an example of which model? : LOWESS model

*You are correct! The answer “*LOWESS model*” is correct because the trend line shows a slight curve that is almost straight.*



A model that defines a person's height as the average of their parents’ heights can be written as (blank). : one half the height of father plus one half the height of mother

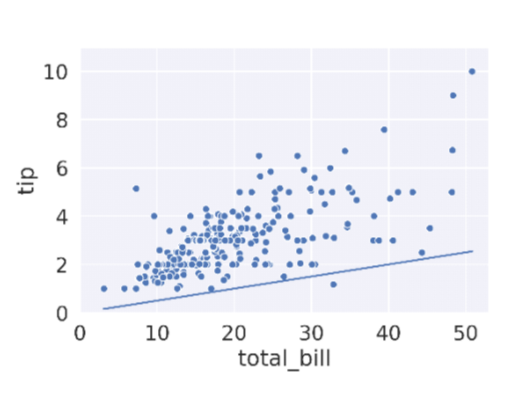
*You are correct! The answer “*one half the height of father plus one half the height of mother*” is correct because the model states it is the average of the parents’ heights.*

If you have six features in a dataset, one of which is the outcome to be predicted, using only one feature for the outcome prediction is a strong model. : False

*You are correct! The answer “*False*” is correct because if a dataset has five features that measure the spread in an outcome, using only a single feature would not be a strong model.*

The graph below shows a trend line that represents 5% of tips of total bills. It is a good prediction model. : False

*You are correct! The answer “*False*” is correct because the trend line is way below the data points, and hence we can say it is underpredicting.*



Which libraries can be used to build a linear regression model in Python? : Plotly, Sklearn

*You are correct! The answers “*Sklearn*” and “*Plotly*” are correct because these Python libraries can be used to build linear regression models.*

What are the steps for building a linear regression model using the Python library Sklearn? : Instantiate a linear model object, Fit the model, Train the model

*You are correct! The answers “*Instantiate a linear model object*”, “*Train the model*”, and “*Fit the model*” are correct because these steps are involved in building a linear regression model using the Python library Sklearn.*

Given a dataset with the outcome variable “tip”, what is the command in the Python library Sklearn to fit a linear regression model? : f.fit(features,tip)

*You are correct! The answer “*f.fit(features,tip)*” is correct because this statement is used in Python library Sklearn to fit a linear regression model.*

The Python library Plotly also needs the object instantiation, fitting, and prediction steps for building a linear regression model. : False

*You are correct! The answer “*False*” is correct because the Python library Plotly does not need these steps for building a linear regression model.*

The name of the constructor used in function px.scatter() in the Python library Plotly to build a linear regression model is “ols.” : False

*You are correct! The answer “*False*” is correct because the name of the constructor is “trendline”, and the value of this constructor used is “ols” to build a linear regression model.*

The function “f.predict()” is used to predict the outcome variable. : True

*You are correct! The answer “*True*” is correct because it predicts the outcome variable values for each instance in the array.*

What is the L2 loss function for a linear regression? : (data[“feature”]-data[“prediction”])\*\*2

*You are correct! The answer “*(data[“feature”]-data[“prediction”])\*\*2*” is correct because this is the formula to calculate the L2 loss.*

The mean squared error of a linear regression model is greater than the mean of the L2 loss. : False

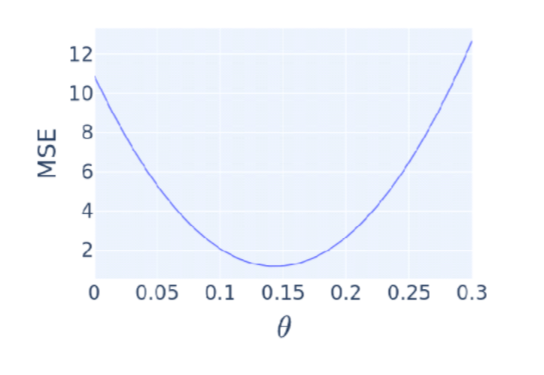
*You are correct! The answer “*False*” is correct because the mean squared error of a linear regression model is equal to the mean of the L2 loss.*

The graph that plots measures of the MSE against the different values of θ shows the most optimum θ to be selected. : True

*You are correct! The answer “*True*” is correct because the most optimal value for theta can be discerned by plotting the MSE against different values of theta.*

What would the most optimal value of θ be in this plot that shows the MSE plotted against the different values of θ? : 0.14

*You are correct! The answer “*0.14*” is correct because this is where the MSE is the lowest.*



What is the function of the Python library SciPy that is used for minimization? : scipy.optimize.minimize()

*You are correct! The answer “*scipy.optimize.minimize()*” is correct because this function is used for minimization.*

The shape of a loss function does not affect optimization or the selection of a minimum. : False

*You are correct! The answer is “*False*” because it is important to pick a loss function that has a nice shape for optimization and a minimum.*

What piece of text shows the failure of the function *“*scipy.optimize.minimize()*”*to find the minimum value*? :* Success: false

*You are correct! The answer “*Success: false*” is correct because this output shows that the function was unable to find the minimum value.*

The absolute value of the difference between prediction and true outcome is known as the MSE. : False

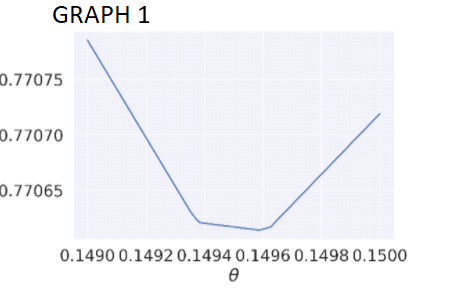
*You are correct! The answer “*False*” is correct because the absolute value of the difference of prediction and true outcome is known as the MAE.*

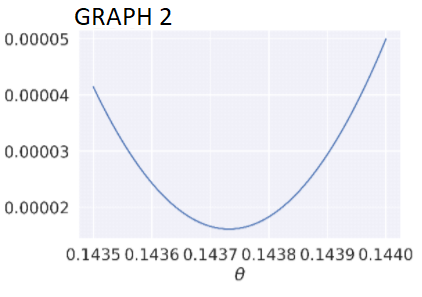
Which of the following loss functions give higher penalties to outliers? : L2

*You are correct! The answer “*L2*” is correct because the L2 loss gives a much higher penalty to outliers than the L1.*

Given these two graphs, which represents the MAE? : GRAPH 1

*You are correct! The answer “*GRAPH 1*” is correct because MAE is piecewise linear. That is, it consists of a bunch of straight-line segments.*





The derivative of the curve of the MSE is discontinuous. : False

*You are correct! The answer “*False*” is correct because the derivative of the curve of the MSE is continuous.*

In multiple linear regression, only one feature can serve as a predictor. : False

*You are correct! The answer “*False*” is correct because in multiple linear regression more than one feature can be a predictor.*

What is the command in Python that retrieves the coefficients of a regression model? : f.coef\_

*You are correct! The answer “*f.coef\_*” is correct because this is the command used in Python to retrieve the coefficients of a regression model.*

Which of the following is an example of multiple linear regression? : Features = data[[“bill”,”size”,”choice”]], Features = data[[“bill”,”size”]]

*You are correct! The answers “*Features = data[[“bill”,”size”]]” and *“*Features = data[[“bill”,”size”,”choice”]]*” are correct because both answers have more than one feature, which is the condition for multiple linear regression.*

Suppose that the outcome variable is “tip”, the features are “Features = data[[“bill”,“size”]]”, and the coefficients for the features are [0.1, 0.3]. What will the linear equation for the outcome variable be? : tip = 0.1\*bill + 0.3\*size

*You are correct! The answer “*tip = 0.1\*bill + 0.3\*size*” is correct because the coefficient is 0.1 for bill is 0.1 and 0.3 for size. Hence, the equation for the tip is correct.*

The output of a linear model is by definition the weighted sum of the features of a given sample. : True

*You are correct! The answer “*True*” is correct because the output of a linear model is the weighted sum of the features of a given sample.*

The approach to create a new set of k features, where k is the number of unique values for the non-numeric feature of interest, is known as (blank). : One-hot encoding

*You are correct! The answer “*One-hot encoding*” is correct because in this approach a new set of k features is created, where k is the number of unique values for the non-numeric feature.*

What is the function in Python that is used for concatenating two dataframes? : concat()

*You are correct! The answer “*concat()*” is correct because the function is used to concatenate pandas objects such as dataframes and series.*

The term used in the generic equation of the linear regression model for the features is σ (“sigma”). : False

*You are correct! The answer “*False*” is correct because the term used for features in the generic equation of the linear regression model is*ϕ*(“phi”).*

A dataframe

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **total\_bill** | **size** | **Thur** | **Fri** | **Sat** | **Sun** |
| **0** | 16.99 | 2 | 0 | 0 | 0 | 1 |
| **1** | 10.34 | 3 | 0 | 0 | 0 | 1 |
| **2** | 21.01 | 3 | 0 | 0 | 0 | 1 |
| **3** | 23.68 | 2 | 0 | 0 | 0 | 1 |
| **4** | 24.59 | 4 | 0 | 0 | 0 | 1 |

Consider the dataframe above. The coefficients for the features are [0.093, 0.187, 0.668, 0.745, 0.621, and 0.732]. What tip would the model predict for a party of three with a $50 total bill eating on a Thursday? : $5.88

*You are correct! The answer “*$5.88*” is correct because with the given information the equation for the model build is “0.093\*50 + 0.187\*3 + 0.668\*1”, which comes out to be 5.88.*

A dataframe

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **total\_bill** | **size** | **Thur** | **Fri** | **Sat** | **Sun** |
| **0** | 16.99 | 2 | 0 | 0 | 0 | 1 |
| **1** | 10.34 | 3 | 0 | 0 | 0 | 1 |
| **2** | 21.01 | 3 | 0 | 0 | 0 | 1 |
| **3** | 23.68 | 2 | 0 | 0 | 0 | 1 |
| **4** | 24.59 | 4 | 0 | 0 | 0 | 1 |

Consider the dataframe above. The coefficients for the features are [0.093, 0.187, 0.668, 0.745, 0.621, and 0.732].

Given this information, how much does the model predict that each person eating at the table will contribute to the tip (in cents)? : 18

*You are correct! The answer “*18*” is correct because the coefficient weight for the feature “size” is 0.187, so depending on how many people are at the table, they throw in 18 cents for each person at the table.*

 ————— o —————