**Introduction**

Welcome to the module on ‘Linear Regression’.

In this module, you will start off with a quick introduction to three machine learning techniques — regression, classification, and clustering. Then, you will go deeper into one of the most important regression models in machine learning — **Linear Regression**.

You will learn primarily about the following two types of linear regression models:

* Simple linear regression
* Multiple linear regression

**In this session**

You will get a quick introduction on machine learning models. Starting with understanding the basics concepts of simple linear regression, you will build a simple linear regression model in Python at the end of this session.

The broad agenda for this session is as follows:

* Introduction to machine learning
* Supervised and unsupervised learning methods
* The linear regression model
* Residuals
* Residual sum of squares (RSS) and R² (R-squared)
* Simple linear regression in Python

# Introduction to Machine Learning

In the previous course, you learnt data preparation, which is the third step of the CRISP-DM framework. Following the same CRISP-DM framework, now let's move to the fourth and the most interesting stage, called **modeling**.

Modeling uses machine learning algorithms, in which the machine learns from the data just like humans learn from their experiences.

Let’s go deeper into the types of models that come under machine learning.

Question:

Question 1

During credit risk analytics, you need to predict which customers would default on their credit card payments and which customers won’t. What type of model will be used to solve this problem?

Top of Form



Regression



**Classification**

**Feedback :***The output variable, in this case, is categorical in nature, i.e. it would be ‘Default’/‘Non-Default’ or 'Yes/No'. So, this is a classification problem.*

**Correct**Bottom of Form

Question 2

Regression & Classification

In credit risk analytics, let’s assume that you need to predict the average amount defaulted by any customer based on different factors. Now, what type of model will be used for this prediction?

Top of Form



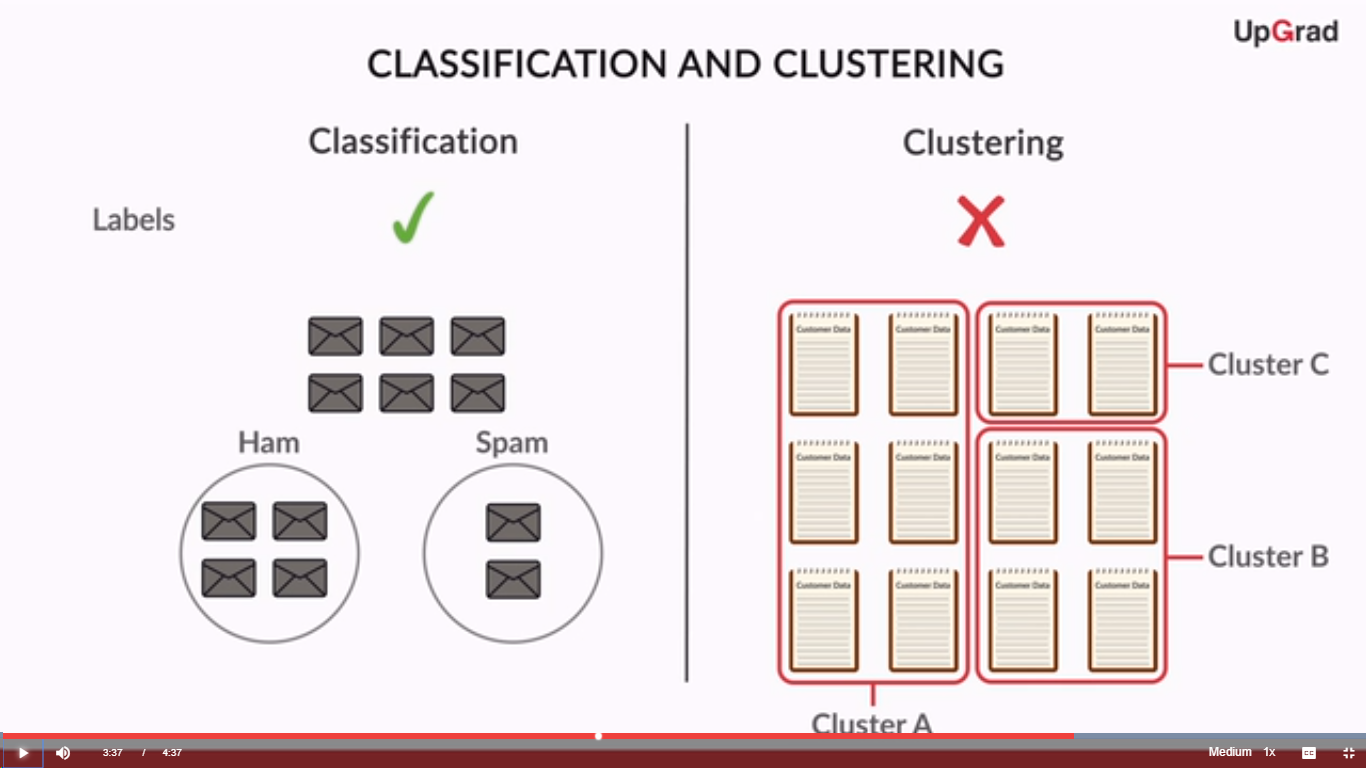
**Regression**

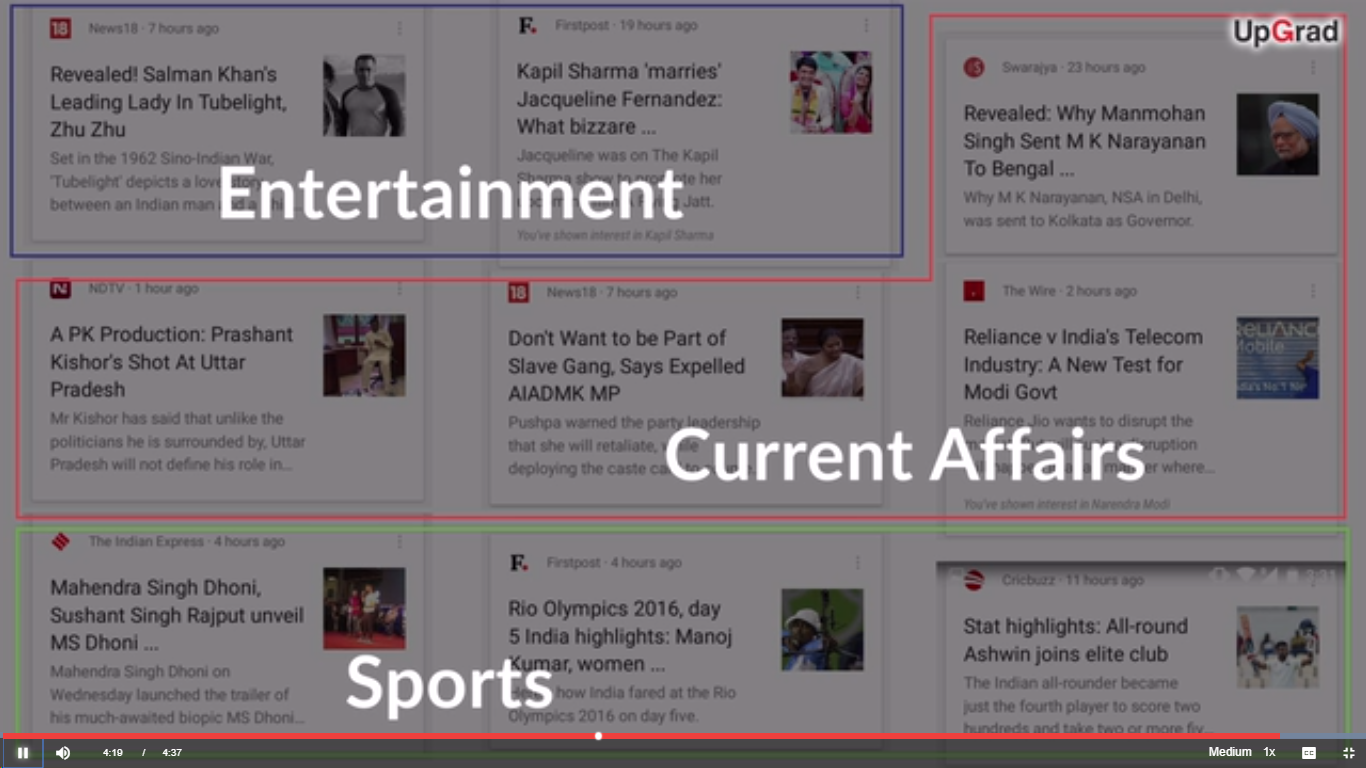
**Feedback :***The average amount defaulted by any customer is a continuous variable. So, a regression model will be used to make the prediction.*

**Correct**



Classification



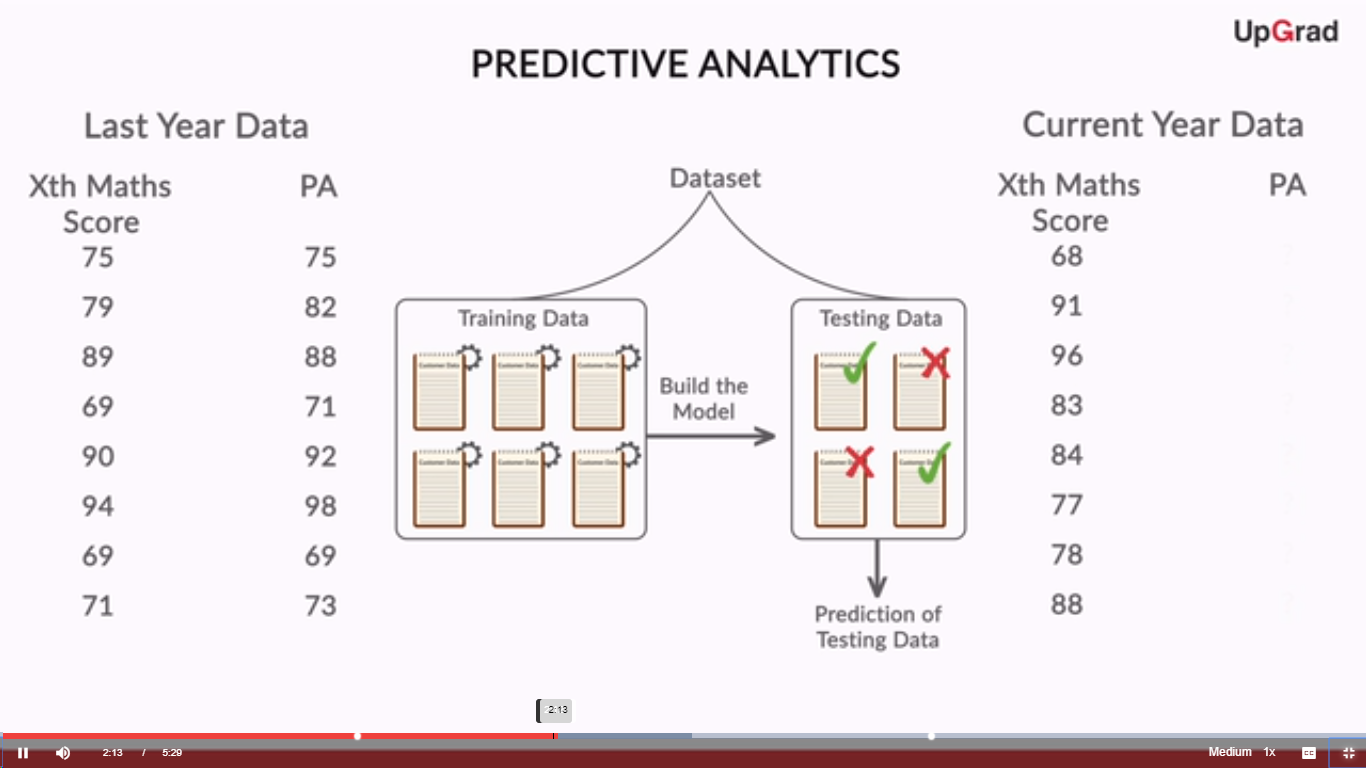


Machine learning models can be classified into the following three types based on the task performed and the nature of the output:

1. **Regression**: The output variable to be predicted is a **continuous variable**, e.g. scores of a student
2. **Classification**: The output variable to be predicted is a **categorical variable**, e.g. classifying incoming emails as spam or ham
3. **Clustering**: **No pre-defined notion of label** allocated to groups/clusters formed, e.g. customer segmentation

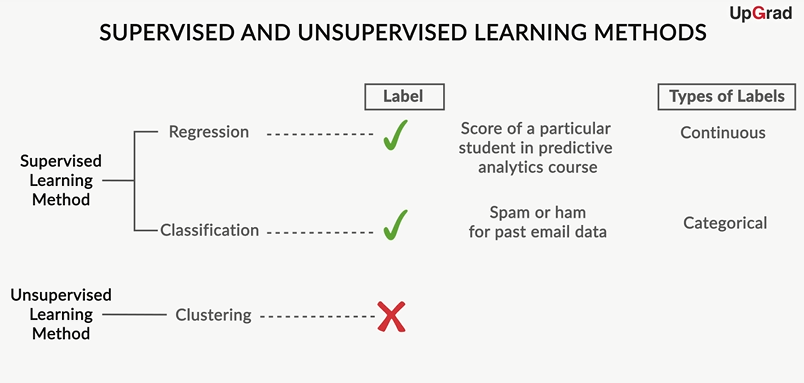
You will get to know about the different types of machine learning models and their classification in detail in the coming modules.

Let’s now learn about supervised and unsupervised learning methods.



So you can classify  machine learning models into two broad categories as follows:

1. **Supervised learning methods**
   1. Past data with labels is used for building the model
   2. **Regression**and **classification** algorithms fall under this category
2. **Unsupervised learning methods**
   1. No pre-defined labels are assigned to past data
   2. **Clustering** algorithms fall under this category

 **Figure 1 - Supervised and Unsupervised Learning Methods**

Mentioned below are some of the problems which are addressed either by a supervised or unsupervised learning algorithm. For each problem, identify if the task can be addressed by a supervised learning algorithm or an unsupervised learning algorithm. Assume an appropriate data set is available for your algorithm to learn from.

Questions:

**Supervised and Unsupervised Learning**

Identify what type of problem this is.

You have the past data of two cricket teams on the performance of the teams based on different parameters and the match results. You have to predict which team will win.

Top of Form



**Supervised learning**

**Feedback :***You have the data of the past two years to train your model on. Since you know the results of different games based on different performance parameters, it would be a supervised learning problem — more specifically, a classification problem since your output variable (i.e. the name of the team) is categorical.*

**Correct**



Unsupervised learning

**Supervised and Unsupervised Learning**

Identify what type of problem this is.

You feed a large collection of spam emails to the learning model to identify the different sub-groups of these spam mails. No labels are presents in the data set.

Top of Form



Supervised learning



**Unsupervised learning**

**Feedback :**This can be addressed using unsupervised learning as there are no labels assigned to your data set and they need to be identified.

**Supervised and Unsupervised Learning**

Identify what type of problem this is.

Consider a large data set of the medical profiles of cancer patients. This data contains no labels for the medical profiles of the cancer patients. The model has to learn whether there might be different groups of such patients for which separate treatments might be tailored.

Top of Form



Supervised learning



**Unsupervised learning**

**Feedback :**This can be addressed using an unsupervised learning algorithm, in which you group patients into different clusters.

# Regression Line

Let’s now go deeper into linear regression.

Question:

Machine Learning Example

You need to predict the total amount of sales, given the marketing spend, from the previous years' data of marketing spend vs sales amount.

What type of machine learning problem is this?

Top of Form



**Regression**

**Feedback :**As the output variable to be predicted (sales amount) is a continuous variable, this would be a regression problem.

**Correct**



Classification

Question:

Independent and Dependent Variable

In credit risk analytics, let’s assume that you need to predict the average amount defaulted by any customer based on different factors such as the credit score of the person, the frequency of using the credit card, the average amount spent during each shopping session, etc.

Can you tell what would be the dependent variable for this regression problem?

Top of Form



Credit score of the person



**Average amount defaulted**

**Feedback :**Dependent variables are the output variables which need to be predicted. For this example, the average amount defaulted by the person is the dependent variable which needs to be predicted.

**Correct**



Frequency of using the credit card



Average amount spent

Question:

Linear Regression

In credit risk analytics, let’s assume that you need to predict the average amount defaulted by any customer based on different factors such as the credit score of the person, the frequency of using the credit card, the average amount spent during each shopping session, etc.

What kind of regression problem is this?

Top of Form



Simple linear regression



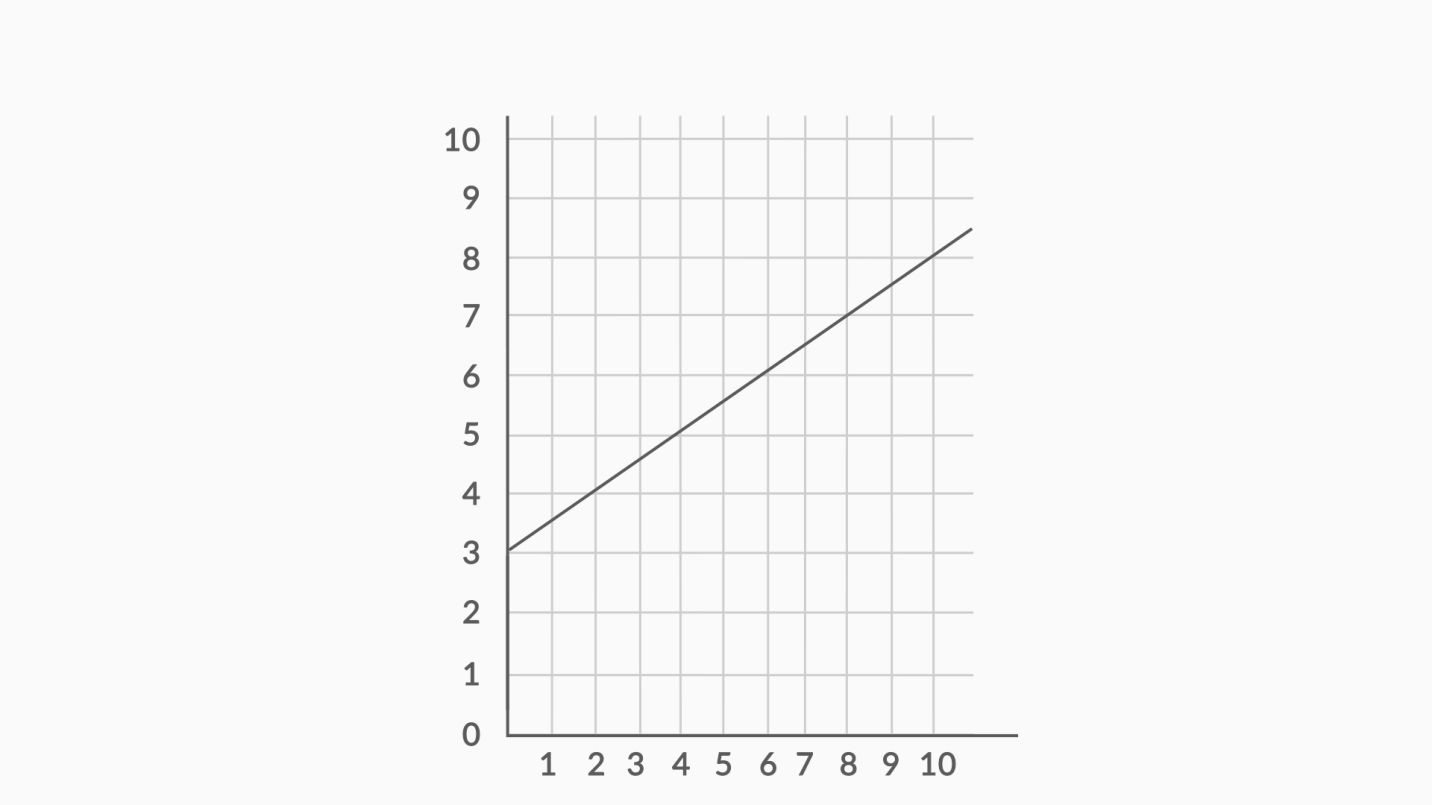
**Multiple linear regression**

**Feedback :**In this case, there are more than 1 independent variables (i.e. the credit score of the person, the frequency of using the credit card, the average amount spent during each shopping session). So, this would be solved using multiple linear regression.

**Correct**

Read [this](https://www.mathsisfun.com/equation_of_line.html) link to revise the physical significance of an equation of a straight line, and also to understand how to find the slope and intercept of a straight line from its graph.

Answer the following questions in the context of the **straight-line plot** given below.

 **Figure 2 - Straight Line graph**

**Questions:**

**Intercept of a Straight Line**

What is the intercept of the given line? Use the graph given above to answer this question.

Top of Form



0



**3**

**Feedback :***The value of y when x = 0 in the given straight line is 3. So, 3 would be the intercept in this case.*

**Correct**



4



½

**Slope of a Straight Line**

What is the slope of the given line? Use the graph given above to answer this question.

Top of Form



**1/2**

**Feedback :***The slope of any straight line can be calculated by (y₂ - y₁)/(x₂ - x₁), where (x₁, y₁) and (x₂, y₂) are any two points through which the given line passes. This line passes (0,3) and (2, 4); so the slope of this line would be (4-3/2-0) = ½.*

**Correct**



1/3



1



2

Bottom of Form

**Equation of a Straight Line**

What would be the equation of the given line?

Top of Form



**Y = X/2 + 3**

**Feedback :***The standard equation of a straight line is y = mx + c, where m is the slope and c is the intercept. In this case, m = ½ and c = 3, so equation would be Y = X/2 + 3.*

**Correct**



Y = 2X + 3



Y = X/3 + ½



Y = 3X + ½

Bottom of Form

Since you now know about the equation of a straight line, let’s look at the general equation of a straight line which is fitted during **simple linear regression**.

Questions:

Coefficients of Regression Line

Which statement explains the physical interpretation of the slope of the regression line Y = β₀ + β₁.X, where Y is the sales amount in crores and X is the marketing spend in lakhs?

Top of Form



For each additional β₁ lakh spent on marketing , the predicted sales increases by 1 crore



**For each additional β₁ lakh spent on marketing , the predicted sales increases by β₀ crore**

Bottom of Form

Top of Form

**For each additional 1 lakh spent on marketing , the predicted sales increases by β₁ crore**

**Feedback :***For this regression line, β₀ is the intercept and β₁ is the slope. For a unit increase in the quantity of X (marketing spend), Y (sales amount) increases by β₁.1 = β₁ units.*

**Correct**



For each additional β₀ lakh spent on marketing , the predicted sales increases by β₁ crore

Bottom of Form

Coefficients of Regression Line

Which statement explains the physical interpretation of the intercept of the regression line Y = β₀ + β₁.X, where Y is the sales amount in crores and X is the marketing spend in lakhs?

Top of Form



**When the amount spent on marketing is Rs. 0, the sales amount = β₀ crore**

**Feedback :**For this regression line, β₀ is the intercept and β₁ is the slope. The value of Y (sales amount) is β₀ crore when the value of X (marketing spend) is Rs. 0.

**Correct**



When the amount spent on marketing is Rs. 0, the sales amount = β₁ crore



When the amount spent on marketing is Rs. β₀ lakh, the sales amount = 0 crore



When the amount spent on marketing is Rs. β₁ lakh, the sales amount = 0 crore

Bottom of Form

Bottom of Form

A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

The independent variable is also known as the **predictor variable**. And the dependent variables are also known as the **output variables**.

# Best Fit Line

In regression, there is a notion of a best-fit line — the line which fits the given scatter-plot in the best way. Let’s look at how you can define the notion of a best-fit line.

Questions:

What will be the residual for x = 6?

Top of Form



Feedback :For x = 2, the actual value of y is 4 and the predicted value of y is 6. So, the value of residual would be = actual value of y - predicted value of y = 4 - 6 = -2.

Marketing and Sales

If ei=yi−yipredei=yi−yipred, what would be the correct expression for e\_{i} in terms of x\_{i},y\_{i}, β₀ & β₁?

Top of Form



​ei=yi−β0+β1xiei=yi−β0+β1xi​



**​**ei=yi−β0−β1xiei=yi−β0−β1xi**​**

**Feedback :**For this regression line y = β₀ + β₁x, β₀, where y is the predicted value of the dependent variable. So substituting β₀ + β₁x in-place of yapredyapred gives this equation of eiei.

**Correct**



​ei=yi+β0+β1xiei=yi+β0+β1xi​



​ei=β0+β1xi−yipredei=β0+β1xi−yipred​

Questions:

**Least Squares Regression Line**

The coefficients of the least squares regression line are determined by the Ordinary Least Squares method — which basically means minimising the sum of the squares of the:

Top of Form



x-coordinates



y-coordinates of actual data



y-coordinates of predicted data



**y-coordinates of actual data - y-coordinates of predicted data**

**Feedback :**The Ordinary Least Squares method has the criterion of the minimisation of the sum of squares of residuals. Residuals are defined as the difference between the y-coordinates of actual data and the y-coordinates of predicted data.

**Best Fit Regression Line**

What is the main criterion used to determine the best-fitting regression line?

Top of Form



The line that goes through the most number of points



The line that has an equal number of points above it or below it



**The line that minimises the sum of squares of distances of points from the regression line**

**Feedback :**

*The criterion is given by the Ordinary Least Squares (OLS) method, which states that the sum of squares of residuals should be minimum. This is explained by this option.*

**Correct**



Either B or C (they are same criterion)

Since now you know that the best-fit line is obtained by minimising a quantity called Residual Sum of Squares (RSS), this is the best time to be introduced to what is known as the **cost function**.

You can learn about the cost function and ways to optimise the cost function (minimisation or maximisation) from [this](https://learn.upgrad.com/v/course/117/session/11510/segment/56853) cost function session given in the optional module.

Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel.

Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel.

Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel. Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel.

Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel.

Question:

Coefficients of Regression Equation

The equation of the regression line here is: y = 0.0528x + 3.3525

What will be the value of β₁ for this regression line?

Top of Form



**0.0528**

**Feedback :**Any general regression equation is represented by y = β₀ + β₁x. So, β₁ (slope of the regression line) is 0.0528.

**Correct**



3.3525



0



1

You can download the Excel file from below and look at the analysis performed by the professor.

# Strength of Simple Linear Regression

After determining the best fit line, there are a few critical questions you need to answer, such as:

1. How well does the best-fit line represent the scatter-plot?
2. How well does the best-fit line predict the new data?

Are the above questions answered well by the RSS? Attempt these two questions on your own before getting the answers from the professor in the following lecture.

**Question:**

**RSS - Residual Sum of Squares**

In the previous example of marketing spend (in lakhs) and sales amount (in crores), let’s assume you get the same data in different units — marketing spend (in lakhs) and sales amount (in dollars). Do you think there will be any change in the value of RSS due to change in units in this case (as compared to the value calculated in the Excel demonstration)?

Top of Form



**Yes, value of RSS would change because units are changing.**

**Feedback :***The RSS for any regression line is given by this expression:*∑(yi−yipred)2∑(yi−yipred)2*. RSS is the sum of the squared difference between the actual and the predicted values, and its value will change if the units change since it has units of ​*y2y2*​. For example, (140 rupees - 70 rupees)^2 = 4900, whereas (2 USD - 1 USD)^2 = 1. So value of RSS is different in both the cases because of different units.*

**Correct**



No, value won’t change



Can’t say

You can play with the interactive graphic below and look at how the position of the regression line and the values of RSS and TSS change with a change in the values of β₀ and β₁.

Let’s go back to our Excel demonstration and see how you can find out the TSS and then R² for the same example for which we had performed regression.

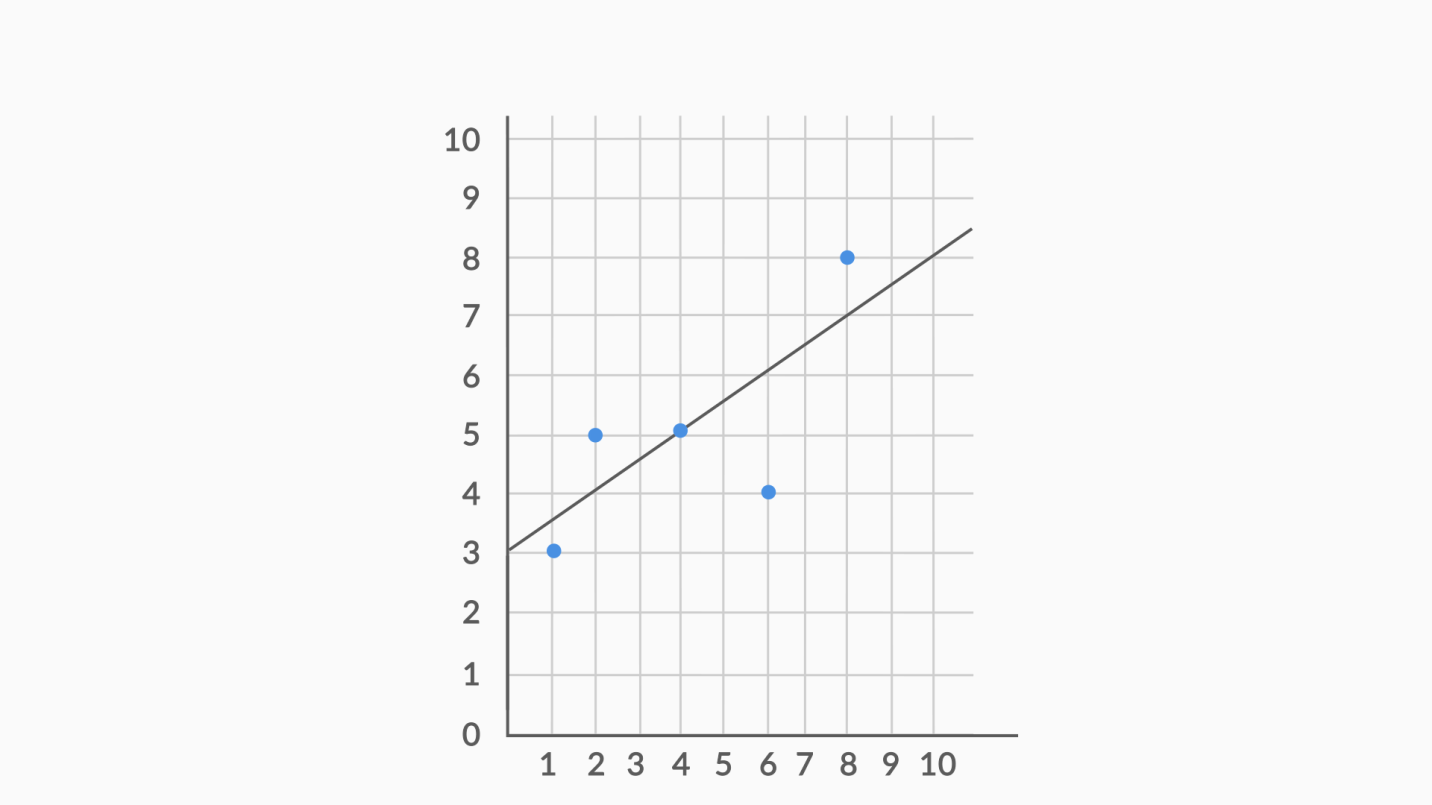
You can download the Excel sheet used in the demonstration for your reference.

**Comprehension**

The plot below represents a scatter-plot of 2 variables X and Y, with the Y variable being dependent on X. Let’s assume the line with the equation Y = X/2 + 3 plotted in the graph represents the best-fit line. This is the same line whose equation you found earlier.

You can find the value of the residual for each point, e.g. for x = 2, the residual would be 5 - 4 = 1.

Answer the following questions in order to consolidate your learning about RSS and R².

 **Figure 3 - Scatter Plot and Regression Line**

**Residual Sum of Squares (RSS)**

Find the value of RSS for this regression line.

Top of Form



0.25



**6.25**

**Feedback :***The residuals for all 5 points are -0.5, 1, 0, -2, 1. The sum of squares of all 5 residuals would be 0.25 + 1 + 0 + 4 + 1 = 6.25*

**Correct**



6.5



-0.5

**Total Sum of Errors (TSS)**

Find the value of TSS for this regression line.

Top of Form



**11.5**

7.5



0



**14**

**Feedback :**The average of y-value for all data points (3 + 5 + 5 + 4 + 8)/5 = 25/5 = 5. So y−¯yy−y¯term for each data point would be -2, 0, 0, -1, 3. So, the squared sum of these terms would be 4 + 0 + 0 + 1 + 9 = 14.

**R²**

The RSS for this example comes out to be 6.25 and the TSS comes out to be 14.

What would be the R² for this regression line?

Top of Form



1 - (14/6.25)



(1 - 14)/6.25



**1 - (6.25/14)**

**Feedback :***R² value is given by 1 - (RSS / TSS). So, in this case, R² value would be 1 - (6.25 / 14).*

**Correct**



(1 - 6.25)/14

Apart from R², there is one more quantity named RSE (Residual Square Error) which is linked to RSS. Let’s see what that is.

In the next segment, you will learn how to build linear regression models in python.

# Simple Linear Regression in Python

Now let’s look at how to build a simple linear regression model in Python.

Please download the data set and the commented Python code from below and run it in the Jupyter Notebook as you go along with the video.

After importing and understanding the data set, let's perform Simple Linear Regression using SKLearn library in python.

We will use the following code:

**from** **sklearn.cross\_validation** **import** train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, train\_size=**0.7** , random\_state=**100**)

We have performed Linear Regression on our advertising dataset, now let's predict using the model we have built.

In the next lecture, we will learn how to predict the labels of the test (unseen by the model) data using the fitted model.

To confirm that we predicted well we need to evaluate our model, let's go through the next lecture and learn how to evaluate the model.

So, to summarise, we learnt implementing Linear Regression on a dataset with a single feature. Also, we understood how to perform linear regression in Python.

**The generic process to build a model in sklearn can be divided into the following steps:**

* **Create X\_train, y\_train, X\_test and Y\_test**

# 1. Create the datasets X\_train, y\_train, X\_test and y\_test

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, train\_size=**0.7** , random\_state=**100**)

* **Create or instantiate an object, e.g. lr = LinearRegression**

# 2. Create (or instantiate) an object of the model you want to build, e.g.

lr = LinearRegression()

* **Fit the model using the object, e.g.  lr.fit(X\_train, y\_train)**

# 3. Fit the model using the training data

lr.fit(X\_train, y\_train)

* **Predict the labels of X\_test, e.g. lr.predict(X\_test)**

# 4. Predict the labels using the test data X\_test

y\_pred = lr.predict(X\_test)

* **Evaluate the model by comparing the predictions with the actual labels**

# 5. Evaluate the model using an appropriate metric by comparing y\_test and y\_predicted

r\_squared = r2\_score(y\_test, y\_pred)

# Coding Practice - Simple Linear Regression

You will now practise building a linear regression model in python.

Many coding exercises from this course onwards will comprise of building ML models, i.e. training a model using some data and making predictions on (unseen) test data.

Since this is the first time you will attempt an ML type question, this video will help you understand how to navigate through these type of questions and explain the procedure briefly.

In the next session, you will study Multiple Linear Regression, i.e. regression with multiple independent variables as the predictors.

# Summary

Here's a brief summary of what you learned in this session:

1. Machine learning models can be classified into following two categories on the basis of learning algorithm:
   * **Supervised learning method:**Past data with labels is available to build the model
     + **Regression:**The output variable is continuous in nature
     + **Classification:**The output variable is categorical in nature
   * **Unsupervised learning method:**Past data with labels are not available
     + **Clustering:** No pre-defined notion of labels is there
2. Past data set is divided into two parts during supervised learning method:
   * **Training data**is used for the model to learn during modelling
   * **Testing data**is used by the trained model for prediction and model evaluation
3. Linear regression models can be classified into two types depending upon the number of independent variables:
   * **Simple linear regression:** When the number of independent variables is 1
   * **Multiple linear regression:** When the number of independent variables is more than 1
4. The equation of the best fit regression line Y = β₀ + β₁X can be found by minimising the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:
   * **Differentiation**
   * **Gradient descent method**
5. The strength of a linear regression model is mainly explained by R²,  whereR² = 1 - (RSS / TSS)
   * **RSS:** Residual Sum of Squares
   * **TSS:** Total Sum of Squares

**Machine Learning Models**

A Singapore-based startup Healin launched an app called JustShakeIt that enables a user to send an emergency alert to emergency contacts and/or caregivers simply by shaking the phone with one hand. The program uses a machine learning algorithm to distinguish between actual emergency shakes and everyday jostling, using data with labels to distinguish between everyday jostling and emergency shaking.

What kind of problem is this?

Top of Form



Supervised learning - Regression



**Supervised learning - Classification**

**Feedback :***The algorithm has to distinguish between actual emergency shakes and everyday jostling. Here, your output variable has pre-defined labels (shake/jostle), which are categorical in nature. So, this is a supervised learning - classification problem.*

**Correct**



Unsupervised learning - Clustering



Can’t say with this information

**Regression in Machine Learning**

Select all the tasks where **linear regression** algorithm can be applied.

More than 1 option can be correct.

Top of Form



**You have a data set of BMI (body-mass index) and fat percentage of customers of a fitness centre. Now, the fitness centre wants to predict the fat percentage of a new customer given his BMI.**

**Feedback :**

Here, the output variable (dependent variable which is to be predicted) is fat percentage, which is a numeric variable. So this is a regression task.

**Correct**



You have collected data from a house rental website such as commonfloor.com. The data has the rental prices of apartments and customer ratings as HIGH or LOW. You want to predict the customer rating given the rental price of a new house.



**You want to predict the sales of a retail store based on its size, given the data set of sales of retail stores and their sizes.**

**Feedback :**

Here, the output variable (dependent variable which is to be predicted) is sales, which is a numeric variable. So this is a regression task.

**Correct**



You want to predict whether a customer is likely to leave the telecom network.

**Coefficients of Regression Equation**

The independent variable X from a linear regression is measured in miles. If you convert it to kilometres (keeping the unit of the dependent Y variable same), how would the slope coefficient change? (1 mile = 1.6 km)

Top of Form



It would get multiplied by 1.6



**It would get divided by 1.6**

**Feedback :**In the linear regression equation, X gets multiplied by 1.6 with no change in Y. So slope will be divided by 1.6.

**Correct**



**It would remain the same**

**Strength of Regression Model**

You want to find out the linear relation between the number of X-ray machines purchased at one time and the cost per machine. The following data has been obtained on the same:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X: Number of machines purchased | 1 | 3 | 6 | 10 | 15 |
| Y: Cost per widget (in thousand rupees) | 89 | 85 | 79 | 73 | 64 |

Suppose the R² between X and Y is 0.87. Which of the following conclusions you can make from this?

Top of Form



The linear relation between X and Y is weak, and their correlation would be positive



The linear relation between X and Y is weak, and their correlation would be negative



The linear relation between X and Y is strong, and their correlation would be positive



**The linear relation between X and Y is strong, and their correlation would be negative**

**Feedback :**A higher value of R² means a strong linear relation. As Y is decreasing with the increasing value of X, you can conclude that their correlation would be negative.

Bottom of Form

Bottom of Form

Bottom of Form

Bottom of Form

Bottom of Form

Bottom of Form

Bottom of Form

You can download the Excel file from below and look at the analysis performed by the professor. You can download the Excel file from below and look at the analysis performed by the professor. Bottom of Form

Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel. Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel. Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel. Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel. Let’s now see a demonstration of obtaining a best-fit line and finding out the RSS for the marketing spend vs sales example in Excel. Bottom of Form

Bottom of Form

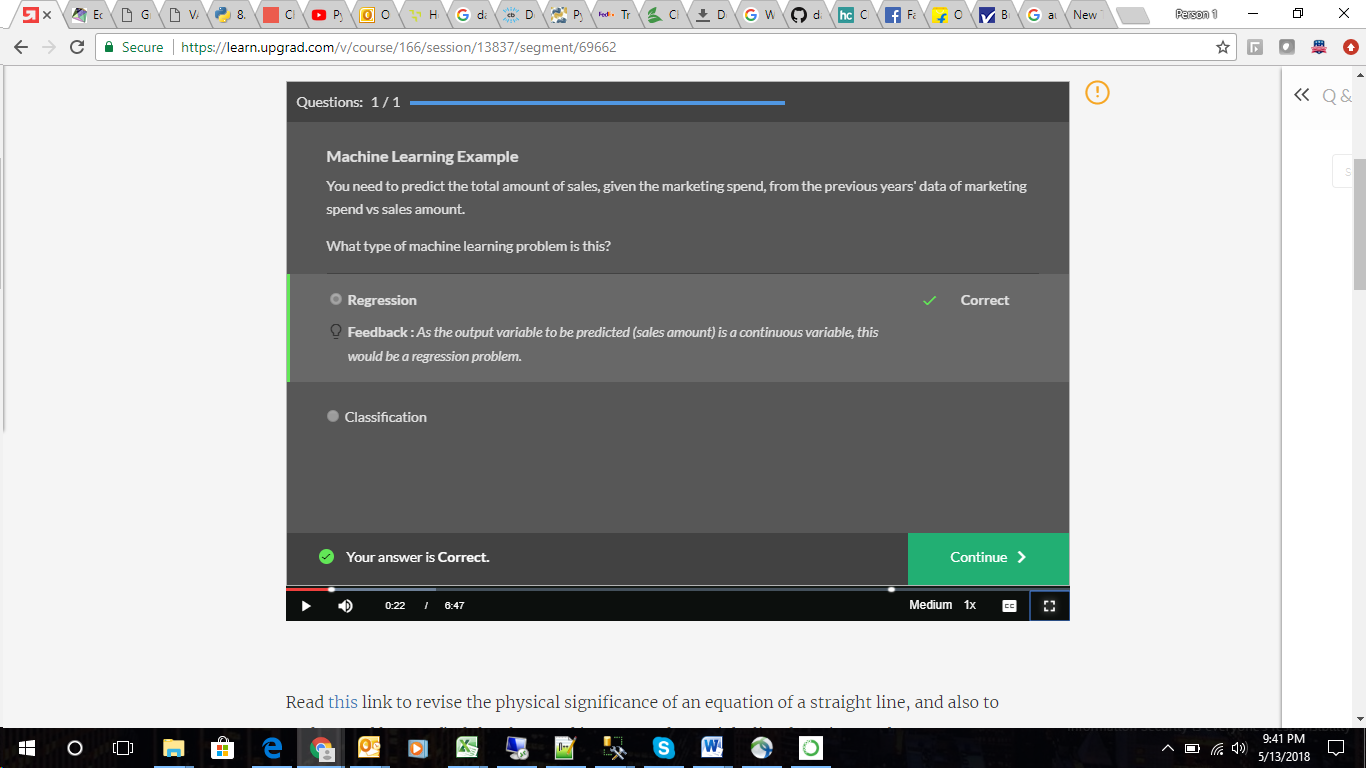
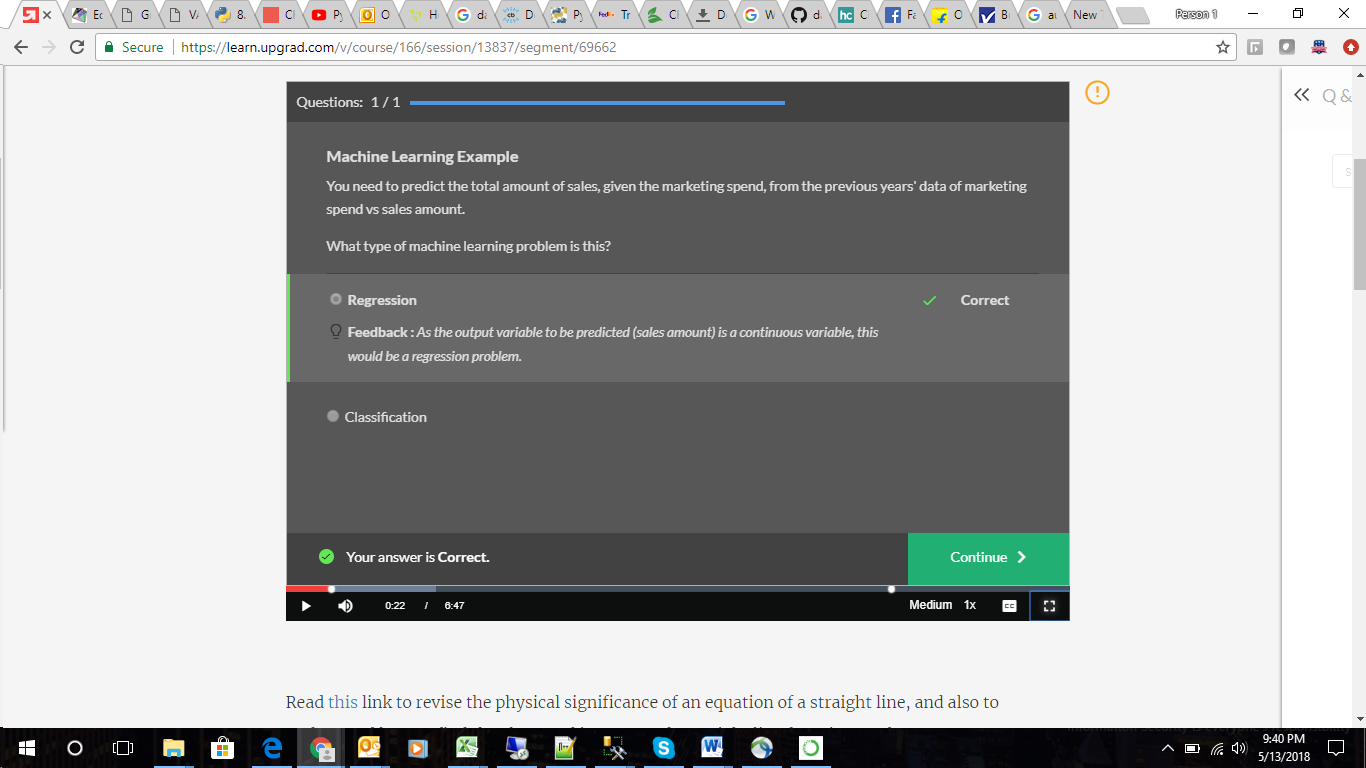
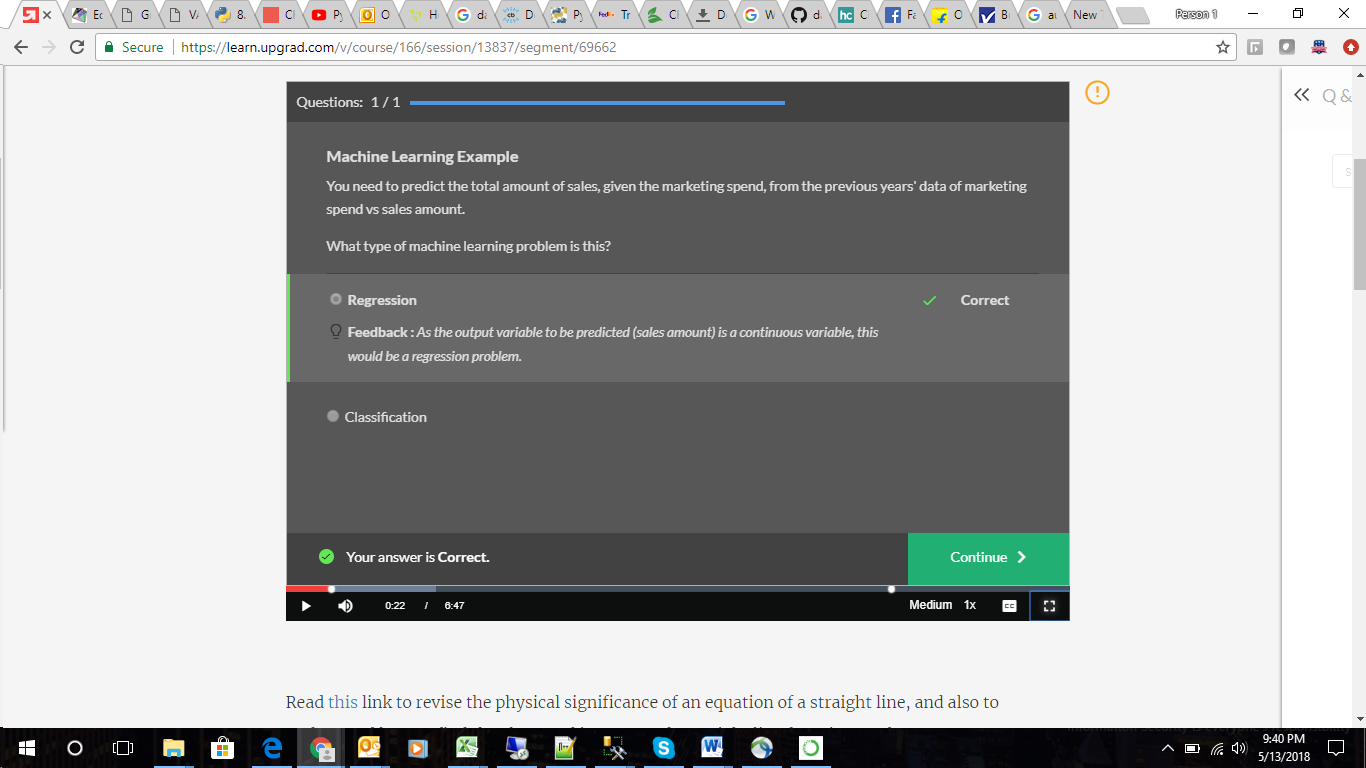
Bottom of Form

Bottom of Form

Bottom of Form

Bottom of Form

Bottom of Form

   Bottom of Form

Bottom of Form

Bottom of Form

Bottom of Form