Welcome to Data Science Online Bootcamp

Week#3_Day#3

dφ

Democratizing Data Science Learning

Learning Objectives

Dependent and Independent Variables

Equation of a Straight Line

Linear Regression

Cost

Dependent and Independent Variables

- So far you’ve been studying input and output/target variables. Commonly, the input variable is known as independent variable and target variable is known as dependent variable.

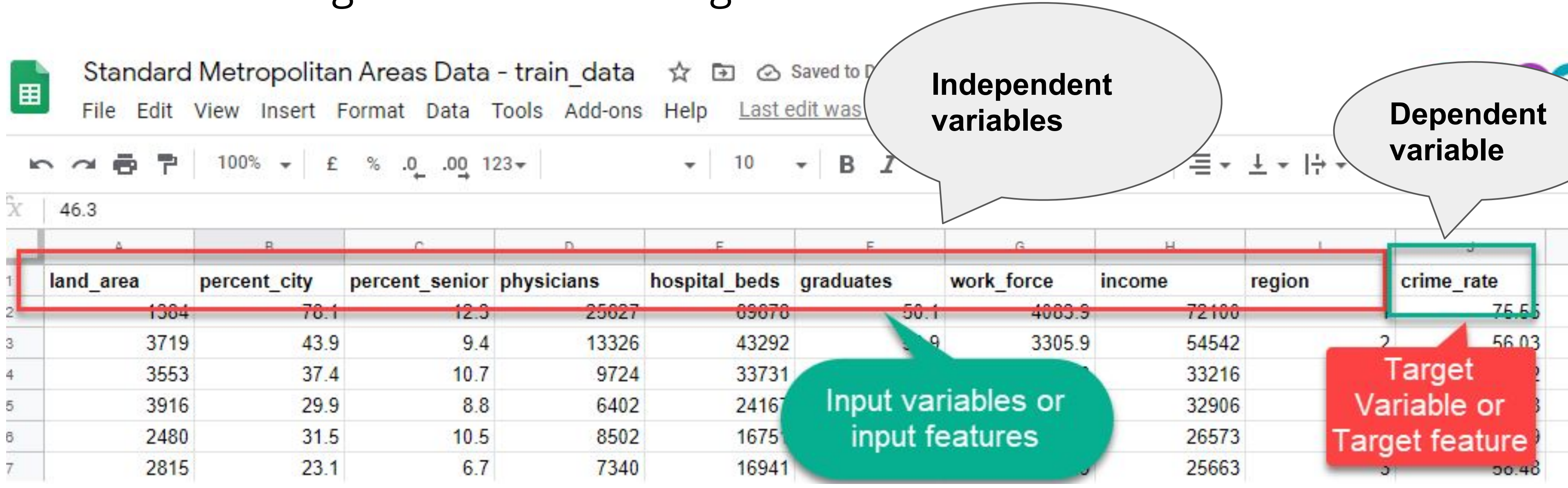
- In nutshell, our target variable is nothing but a dependent variable. Why dependent? Because the values of this variable are dependent on other variables (i.e. input variables)

- And, our input variables are known as independent variables. Here the values of these variables are not dependent on any other variables.

Let’s look at some examples to learn more about them!

Dependent and Independent Variables

- Remember the Standard Metropolitan Areas Data used in previous slides? In that dataset **we might be curious to predict “crime_rate” in future**, so that becomes our target variable **(dependent variable)** and rest of the variables become input variables (**independent variables**) for building a machine learning model.


Standard Metropolitan Areas Data - train_data
Saved to D
File Edit View Insert Format Data Tools Add-ons Help
Last edit was
100%
£ % .0 .00 123
10
B I
Independent variables
Dependent variable
Input variables or input features
Target Variable or Target feature

fx 46.3

	A	B	C	D	E	F	G	H	I	J
1	land_area	percent_city	percent_senior	physicians	hospital_beds	graduates	work_force	income	region	crime_rate
2	1384	78.1	12.3	25627	69678	50.1	4083.9	72100	1	75.55
3	3719	43.9	9.4	13326	43292	[illegible]	3305.9	54542	2	56.03
4	3553	37.4	10.7	9724	33731	[illegible]	[illegible]	33216	[illegible]	[illegible]
5	3916	29.9	8.8	6402	2416	[illegible]	[illegible]	32906	[illegible]	[illegible]
6	2480	31.5	10.5	8502	1675	[illegible]	[illegible]	26573	[illegible]	[illegible]
7	2815	23.1	6.7	7340	16941	[illegible]	[illegible]	25663	3	58.48

Another example

- For example, a scientist wants to see if the brightness of light has any effect on a moth being attracted to the light.

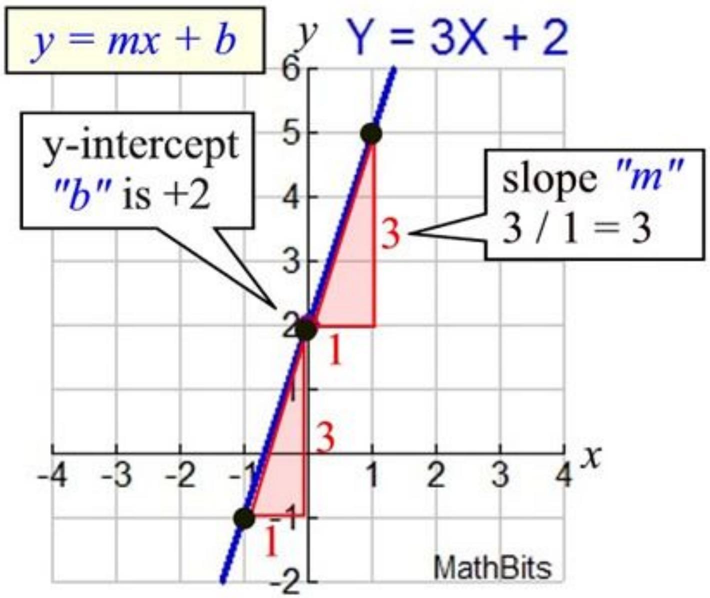
- The brightness of the light is controlled by the scientist. This would be the independent variable.

- How the moth reacts to the different light levels (distance to light source) would be the dependent variable.

Equation of a Straight Line

- In algebra, a linear equation (equation of a straight line) typically takes the form y = mx + b, where m and b are constants, **x** is the **independent variable**, **y** is the **dependent variable**.

- Basically, the value of y is being calculated using x whereas x has no dependence on value of y.



➔ y = how far up
➔ x = how far along
➔ m = Slope or Gradient (how steep the line is)
➔ b = value of y when x=0

- **How do you find "m" and "b"?**

  - b is easy: just see where the line crosses the Y axis.
  - m (the Slope) needs some calculation:

$$m = \frac{\text{Change in Y}}{\text{Change in X}}$$

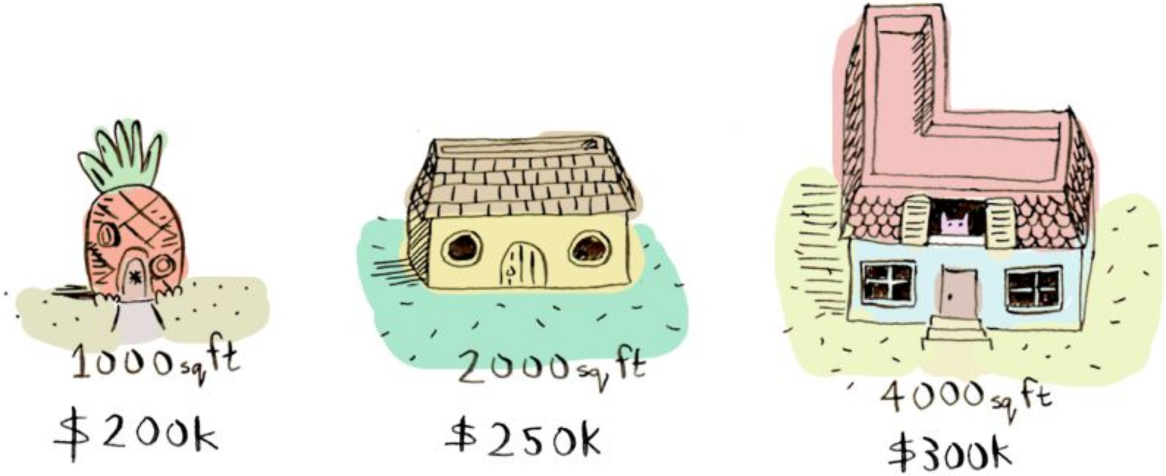
Synonyms Recap

Too many synonyms to memorise? Let me put them all down at one place for better understanding:

- Variables = Features
- Input Variables = Attributes = Predictor = Independent Variables
- Target Variables = Labels = Outcomes = Dependent Variables

What is linear regression? - an example

Suppose you are thinking of selling your home. And, various houses around you with different sizes (area in sq.ft) around you have sold for different prices as listed below:

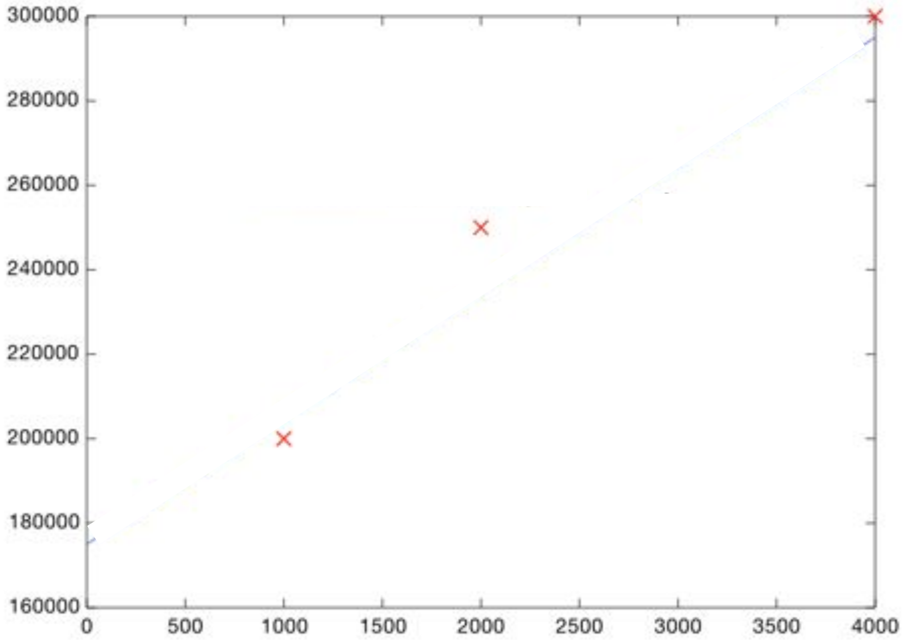


And considering, **your home is 3000 square feet**. How much should you sell it for?

Well! You have to **look at the existing price patterns (data) and predict a price for your home.** This is called **linear regression.**

What is linear regression? - an example

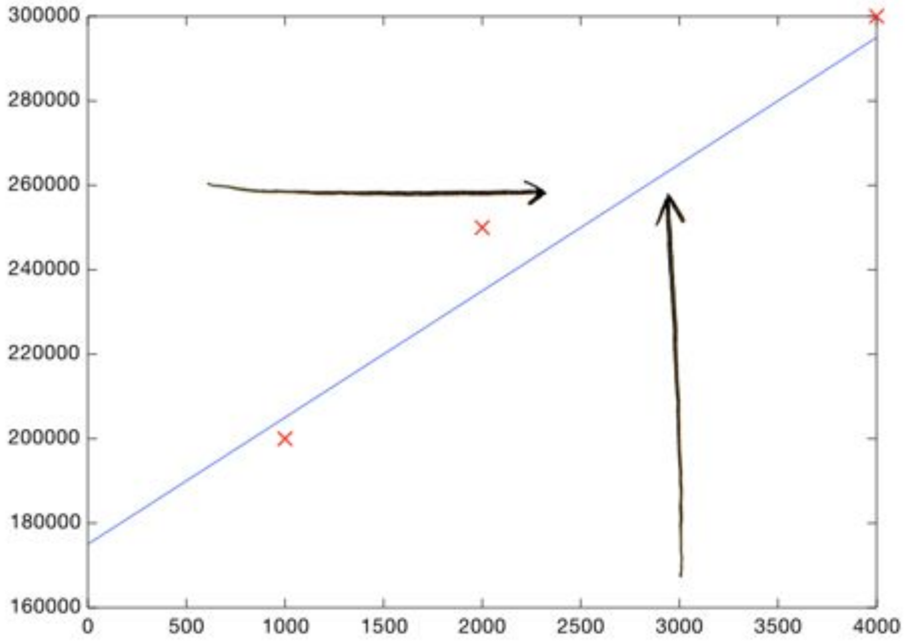
Here's an easy way to do it. Plotting the 3 data points we have so far:



Each point represents one home.

What is linear regression? - an example

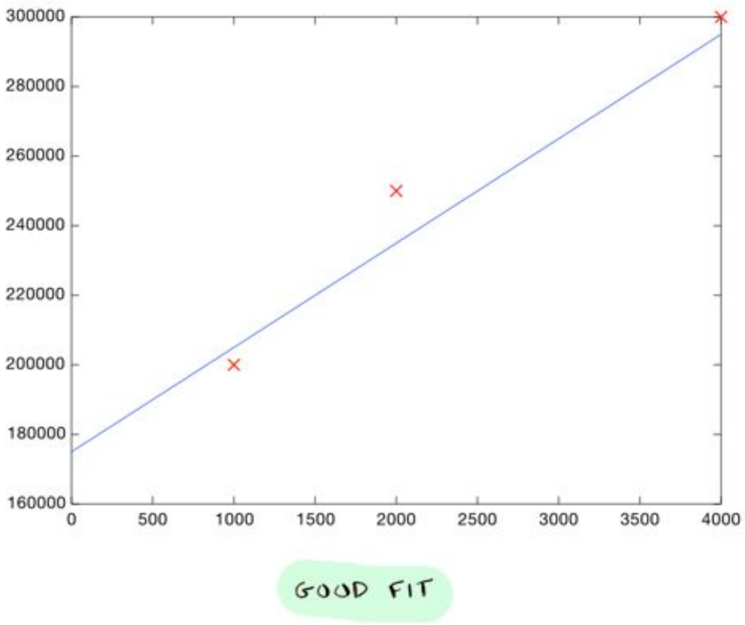
Now you can eyeball it and roughly draw a line that gets pretty close to all of these points. Then look at the price shown by the line, where the square footage is 3000:

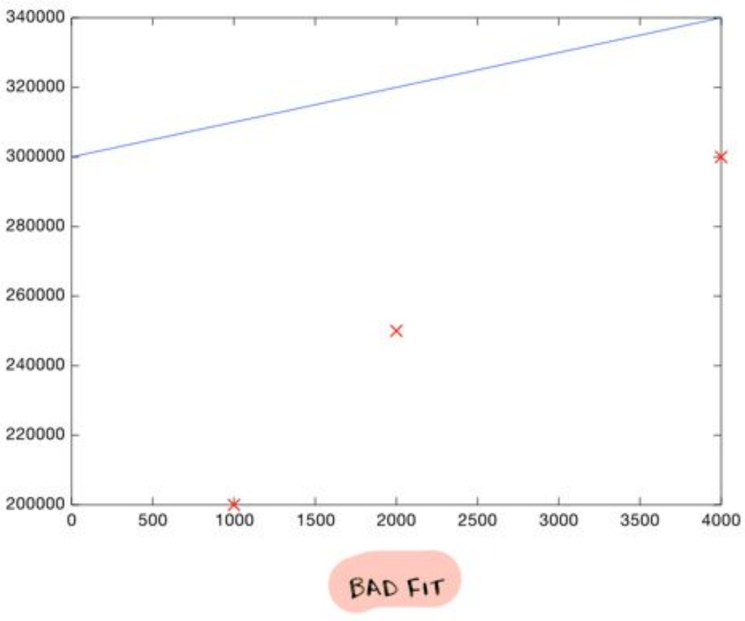


Boom! Your home should sell for $260,000.

What is linear regression? - an example

That's all! You plot your data, make a rough line, and use the line to make predictions. You need to make sure your line fits the data well:





But of course we don't want to roughly make a line, we want to compute the exact line that best "fits" our data. That’s where machine learning comes into play!

What is linear regression?

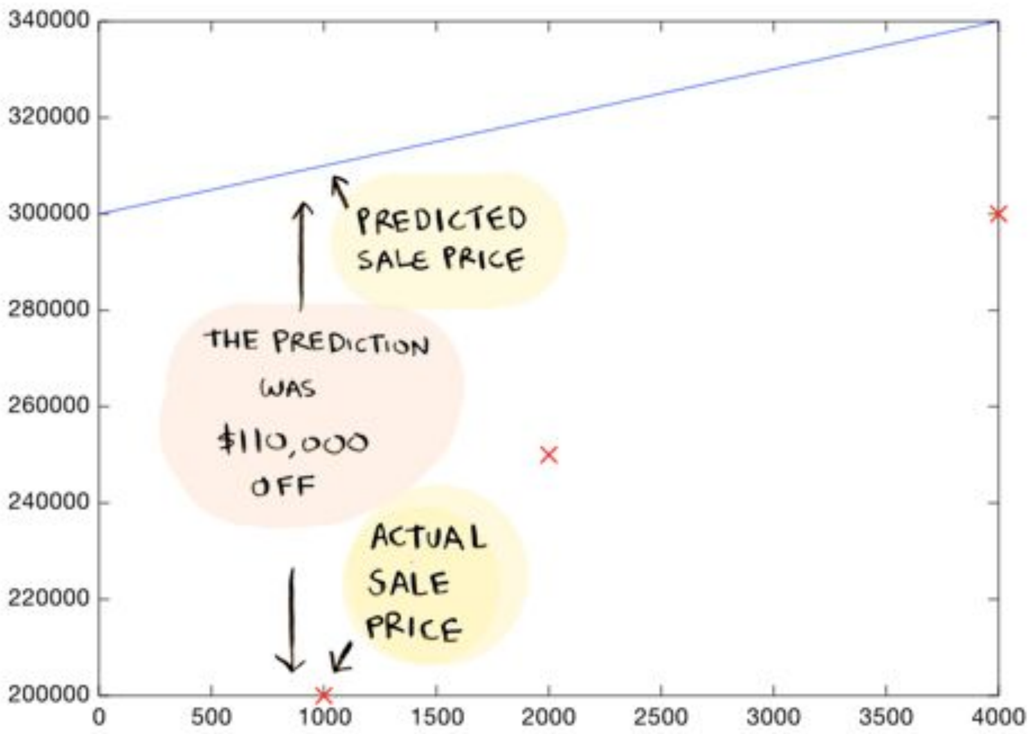
- Linear regression is a linear model i.e. a model that assumes a **linear relationship** (straight-line relationship) between the input variables (x) and the single output variable (y).

- When there is a single input variable (x), the method is referred to as **simple linear regression** or just linear regression. **Eg:** Salary dataset given here. There is only one target variable and one input variable where we are predicting the salary of individual using their years of experience.

- When there are multiple input variables, it is often referred to as **multiple linear regression**. **Eg:** Smart Metropolitan areas data set, we have multiple input variables

Which line is good?

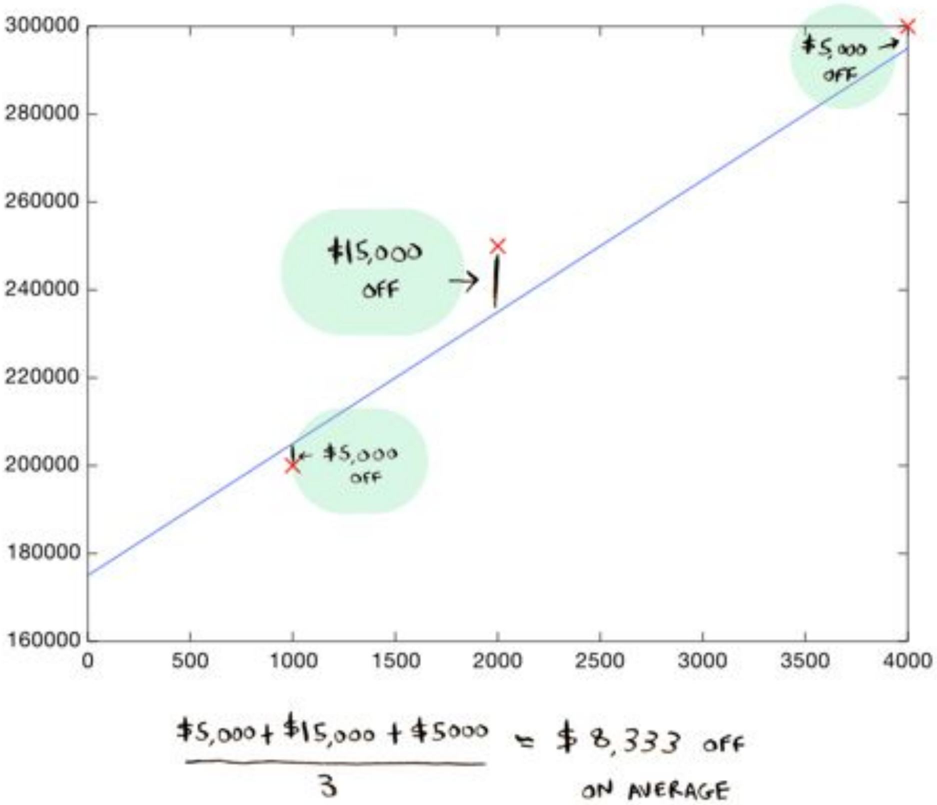
How do you decide what line is good? Here's a bad line:



This above drawn line is way off. For example, according to the line, a 1000 sq foot house should sell for $310,000, whereas we know it actually sold for $200,000.

Which line is good?

Here's a better line:



This line is an average of $8,333 dollars off (adding all the distances and dividing by 3).

This $8,333 is called the **cost** of using this line.

Short-term Objective

What were we doing in the previous 2 examples? We plotted 2 straight lines using the equation: y = mx+b.

If we already have the data points (x1, y1), ..., (xn, yn), it means that our values of x and y remain the same throughout all the lines we plot.

So what remains? What exactly are we changing to plot different lines? Yes, m and b.

Our objective is to find the values of m and b that will best fit this data.

These 2 variables are actually called **hyperparameters.** In machine learning, **a hyperparameter is a parameter whose value is used to control the learning process. And we must always try to find some optimal parameters while building a machine learning model.**

Cost

The **cost** is how far off the line is from the real data. The best line is the one that is the least off from the real data.

To find out what line is the best line (to find the values of m and b), we need to use a **cost function**.

In ML, cost functions are used to estimate how badly models are performing.

Put simply, a cost function is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and y.

Cost function

There is also something called as Cost function that is associated with the analysis, it is slightly mathematical, we will be learning more about it soon!

Meanwhile, let’s run a simple linear regression model.

Notebook for practice

https://github.com/dphi-official/Linear_Regression_Introduction

- Download
- Extract zip file
- Open in Jupyter Notebook or Upload on Google Colab

References

- http://adit.io/posts/2016-02-20-Linear-Regression-in-Pictures.html

Slide Download Link

- You can download the slides here: https://docs.google.com/presentation/d/1Vhgfhjc3Ye90wjgy_RPt-G2ModxN1O9EI4kfOa2SS0M/edit?usp=sharing

That’s it for the day. Thank you!

Feel free to post any queries in the #help channel on Slack