

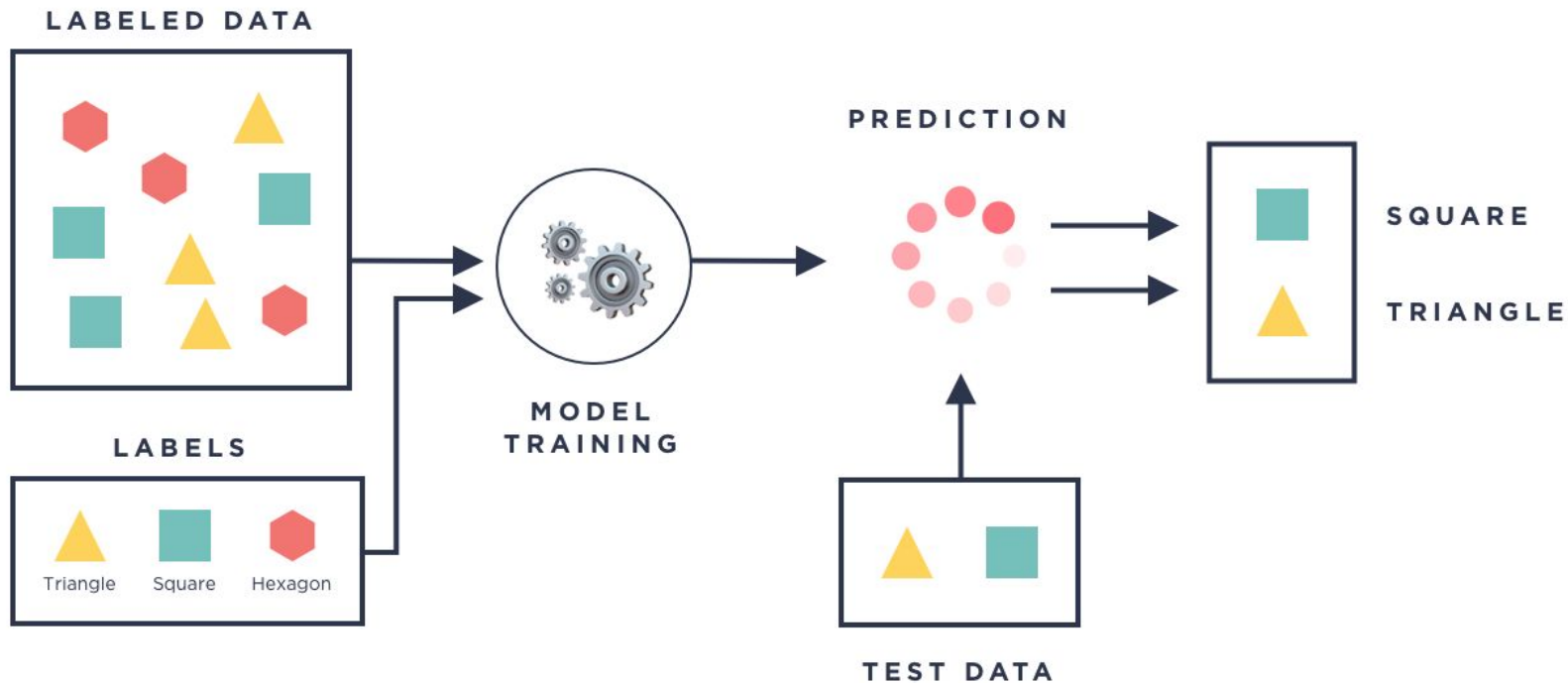


Data Analytics Full-Time Bootcamp

1.06 - Linear Regression



Supervised Learning.



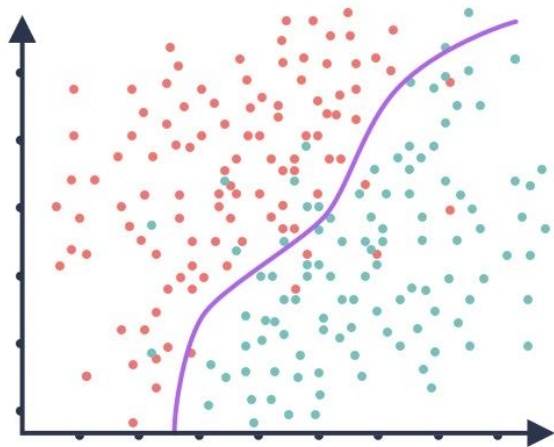
Features

Labels

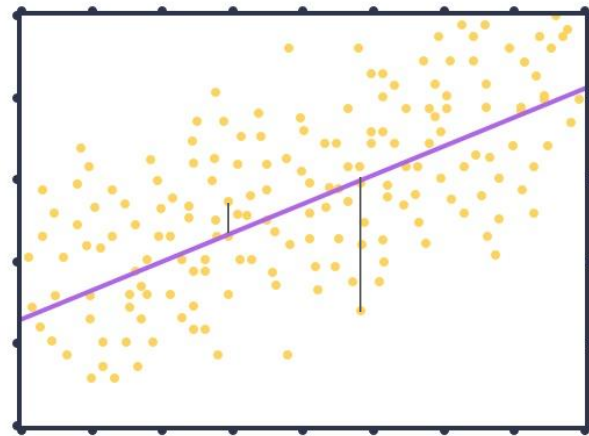


HV1	IC1	IC2	IC3	IC4	IC5	AVGGIFT	TARGET_D
2346	420	446	468	503	14552	15.5	21
497	350	364	357	384	11696	3.08	3
1229	469	502	507	544	17313	7.5	20
325	148	181	171	209	6334	6.7	5
768	174	201	220	249	7802	8.78571429	10
557	211	188	221	205	5550	13	16
2145	474	492	522	554	18340	11.5714286	15
2184	351	376	394	419	16480	12.5	20
1442	369	394	445	488	26462	7.84615385	10
1708	437	586	551	684	29098	9.76923077	20
1054	584	644	652	726	26074	13.5384615	20
1062	486	550	555	584	17908	15.3333333	20
849	457	508	470	519	16386	12.8	25
213	222	273	283	329	12227	5.125	5
574	289	318	315	363	11250	3.55555556	4
2506	449	455	501	517	16302	8.875	50
622	347	378	401	416	15808	15	25
764	272	361	346	424	16257	7.91304348	15
681	335	398	356	419	14011	30.75	51

What is the difference between...



CLASSIFICATION



REGRESSION

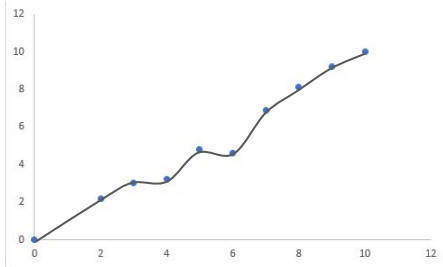
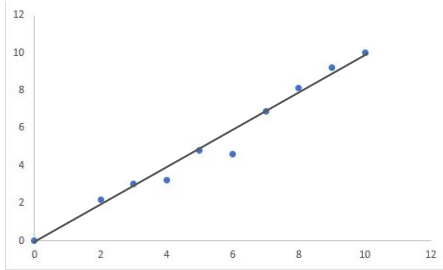




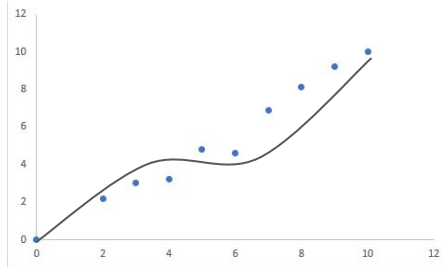
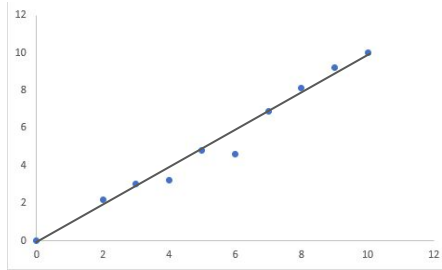
Linear Regression.

Challenges in Regression

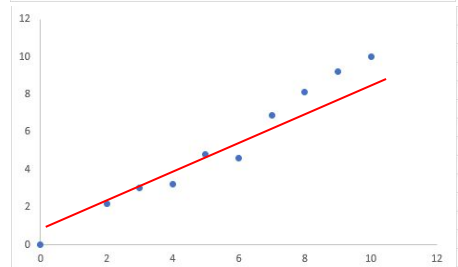
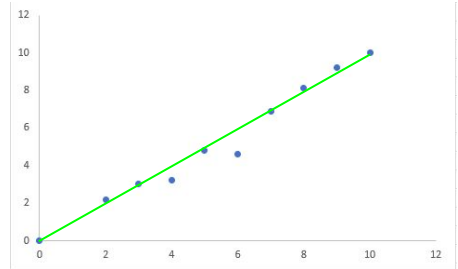
What does it mean to have a “good fit”



What kind of functions do we use to approximate our data?



How do we choose the best among those functions?



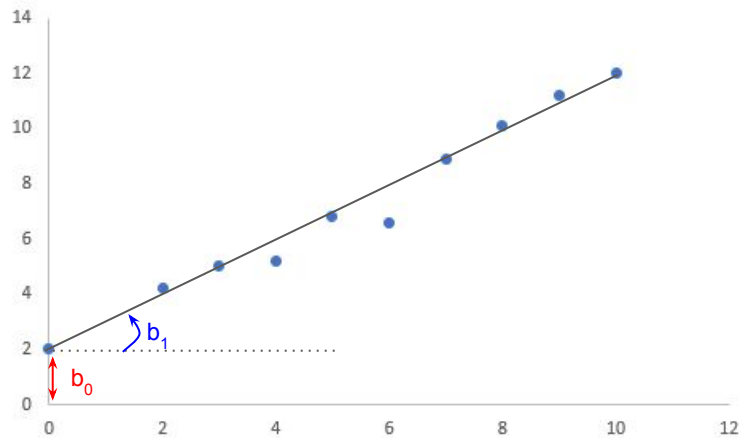
Linear Regression

We approximate our points by a straight line.
Why?

- It's simple
- It's easy to explain
- It works surprisingly well

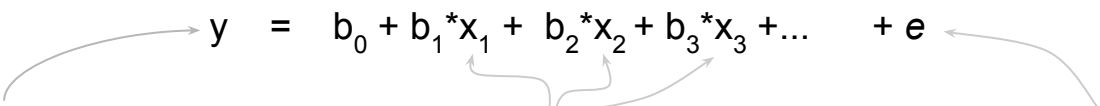
A line can be determined by two parameters

$$y = b_0 + b_1 * x$$



[Go to class script](#)

Linear Regression - Assumptions



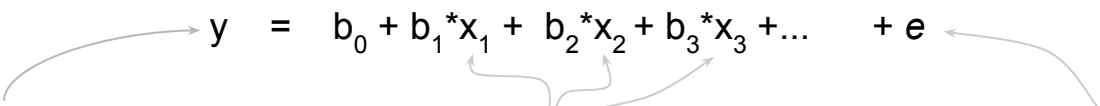
The diagram shows the linear regression equation $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + e$. Arrows point from descriptive labels below to parts of the equation: from 'True value of the label' to y , from 'Values of observed features' to the feature terms $b_1x_1 + b_2x_2 + b_3x_3 + \dots$, and from 'Error of prediction' to $+ e$.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + e$$

True value of the label Values of observed features Error of prediction

- Assumption 1: The label is linear on the features
- Assumption 2: Error term e is well modelled by a Normal distribution
- Assumption 3: Error term e does not change with x
- Assumption 4: No collinearity between features

Linear Regression - Assumptions



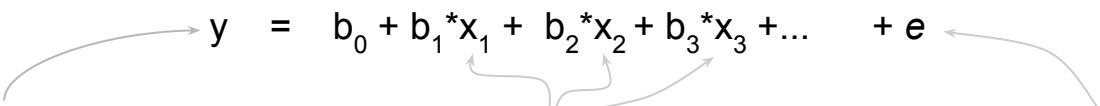
The diagram shows the linear regression equation $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + e$. A curved arrow points from the text 'True value of the label' to the variable y . A bracket under the terms $b_1x_1 + b_2x_2 + b_3x_3 + \dots$ is connected by a line to the text 'Values of observed features'. A curved arrow points from the text 'Error of prediction' to the term $+ e$.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + e$$

True value of the label Values of observed features Error of prediction

- Assumption 1: The label is linear on the features Pretty obvious
- Assumption 2: Error term e is well modelled by a Normal distribution See next slide
- Assumption 3: Error term e does not change with x See next slide
- Assumption 4: No collinearity between features See class script

Linear Regression - Assumptions



The diagram shows the linear regression equation $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + e$. Arrows point from descriptive labels below to parts of the equation: from 'True value of the label' to y , from 'Values of observed features' to the feature terms $b_1x_1 + b_2x_2 + b_3x_3 + \dots$, and from 'Error of prediction' to $+ e$.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + e$$

True value of the label Values of observed features Error of prediction

- Assumption 2: Error term e is well modelled by a Normal distribution
- Assumption 3: Error term e does not change consistently with x

Say your Errors do not follow these assumptions.
Then you are leaving information on the table. Exploit that information and build a better model.