

annotated
version

Machine Learning Course - CS-433

Regression

Sept 17, 2019

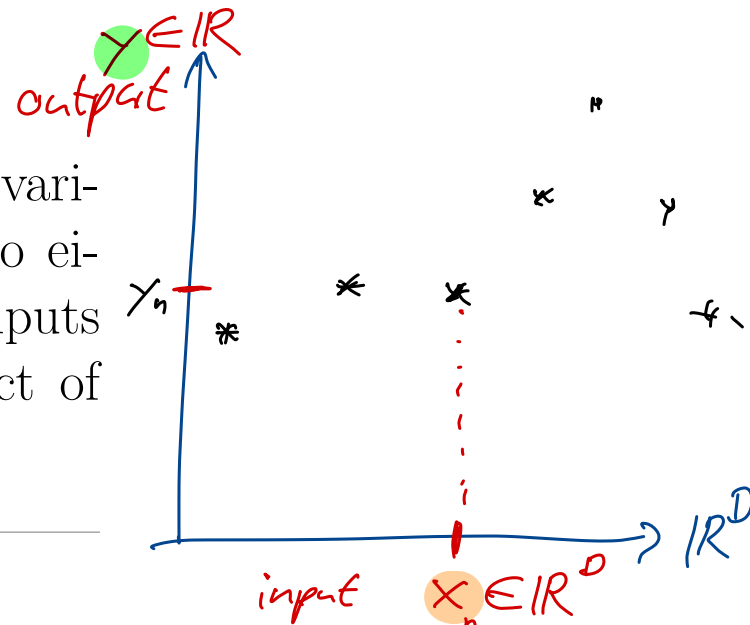
minor changes by Martin Jaggi 2019,2018,2017,2016; ©Mohammad Emtiyaz Khan 2015

Last updated on: September 17, 2019

EPFL

What is regression?

Regression is to relate input variables to the output variable, to either predict outputs for new inputs and/or to understand the effect of the input on the output.



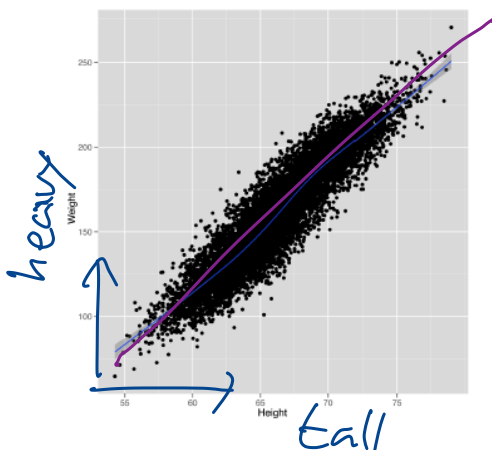
Dataset for regression

In regression, data consists of pairs (\mathbf{x}_n, y_n) , where y_n is the n 'th output and \mathbf{x}_n is a vector of D inputs. The number of pairs N is the data-size and D is the dimensionality.

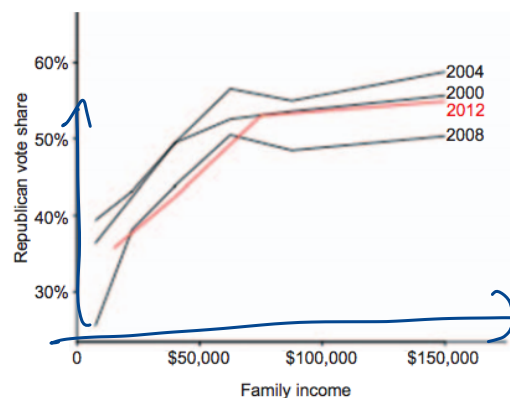
$$\text{Data} = \left\{ \left(\underbrace{\mathbf{x}_n}_{\mathbb{R}^D}, \underbrace{y_n}_{\mathbb{R}} \right)_{n=1}^N \right\}$$

N	x_{11}	x_{12}	\dots	x_{1D}	y_1
	x_{n1}	x_{n2}	\dots	x_{nD}	y_n
	x_{N1}	x_{N2}	\dots	x_{ND}	y_N
	D				

Examples of regression



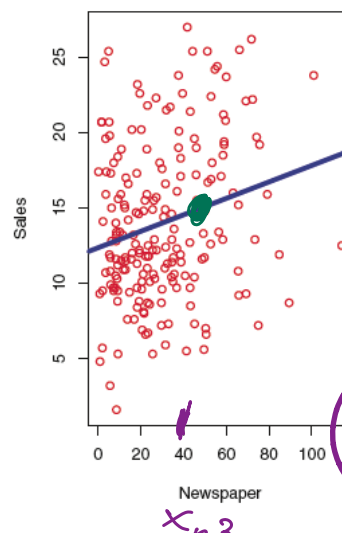
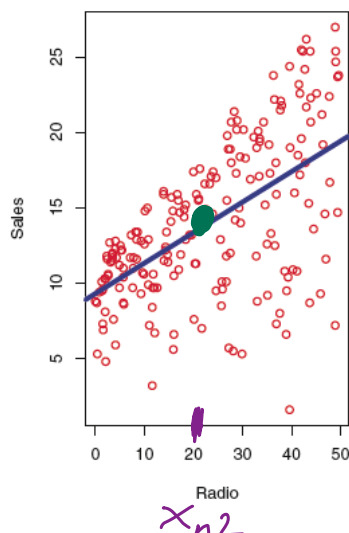
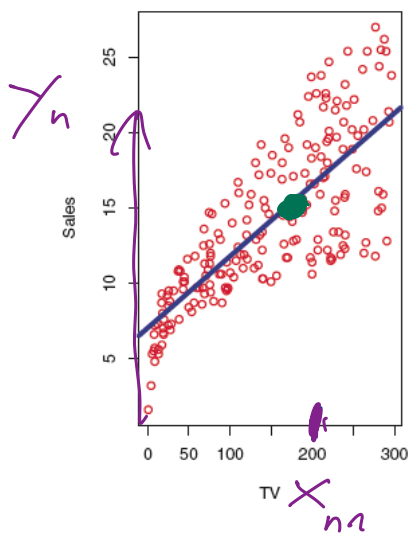
(a) Height is correlated with weight. Taken from “Machine Learning for Hackers”



$D=1$

one-dimensional

(b) Do rich people vote for republicans? Taken from Avi Feller et. al. 2013, Red state/blue state in 2012 elections.



$$D = 3$$

multi-variate regression

$$\mathbf{x}_n = (200, 20, 40)$$

(c) How does advertisement in TV, radio, and newspaper affect sales? Taken from the book "An Introduction to statistical learning"

$$y_n = 15$$

Two goals of regression

In **prediction**, we wish to predict the output for a new input vector, e.g. what is the weight of a person who is 170 cm tall?

In **interpretation**, we wish to understand the effect of inputs on output, e.g. are taller people heavier too?

The regression function

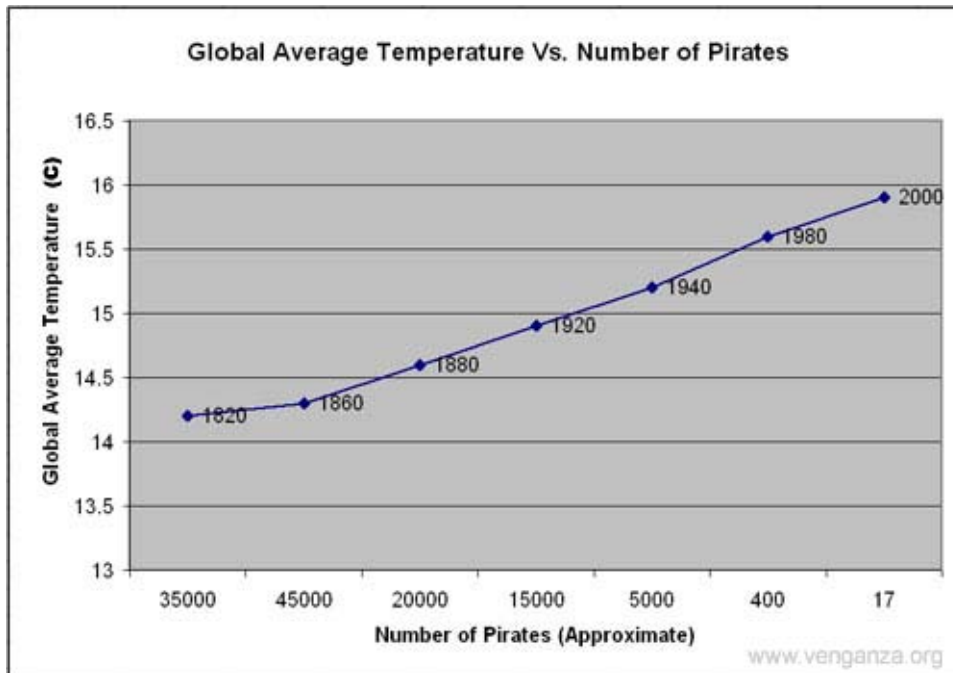
For both the goals, we need to find a function that approximates the output "well enough" given inputs.

$$y_n \approx f(\mathbf{x}_n), \text{ for all } n$$

Additional Notes

Correlation \neq Causation

Regression finds correlation not a causal relationship, so interpret your results with caution.



This image is taken from <http://www.venganza.org/>. You can see many more examples at this page: [Spurious correlations page](#).

Machine Learning Jargon for Regression

Input variables are also known as features, covariates, independent variables, explanatory variables, exogenous variables, predictors, regressors.

Output variables are also known as target, label, response, outcome, dependent variable, endogenous variables, measured variable, regressands.

Prediction vs Interpretation

Some questions to think about: are these prediction tasks or interpretation task?

1. What is the life-expectancy of a person who has been smoking for 10 years?
2. Does smoking cause cancer?
3. When the number of packs a smoker smokes per day doubles, their life span gets cut in half?
4. A massive scale earthquake will occur in California within next 30 years.
5. More than 300 bird species in north America could reduce their habitat by half or more by 2080.

Homework