Linear Regression

Lecture 2
13th of July

Definition

Linear regression models are used to predict the relationship between two variables:

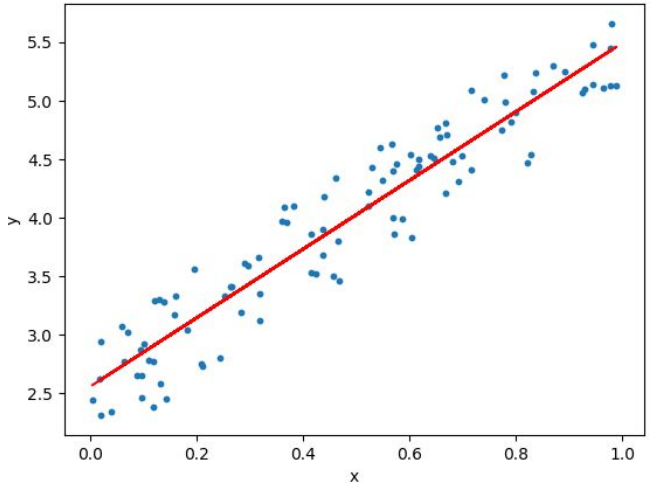
Dependent Variable - factor that is being predicted

Independent Variables - factors that are used to predict the value of the dependent variable

$$Y = aX + b$$

More Definitions

Linear regression involves finding the best-fitting straight line through the points.



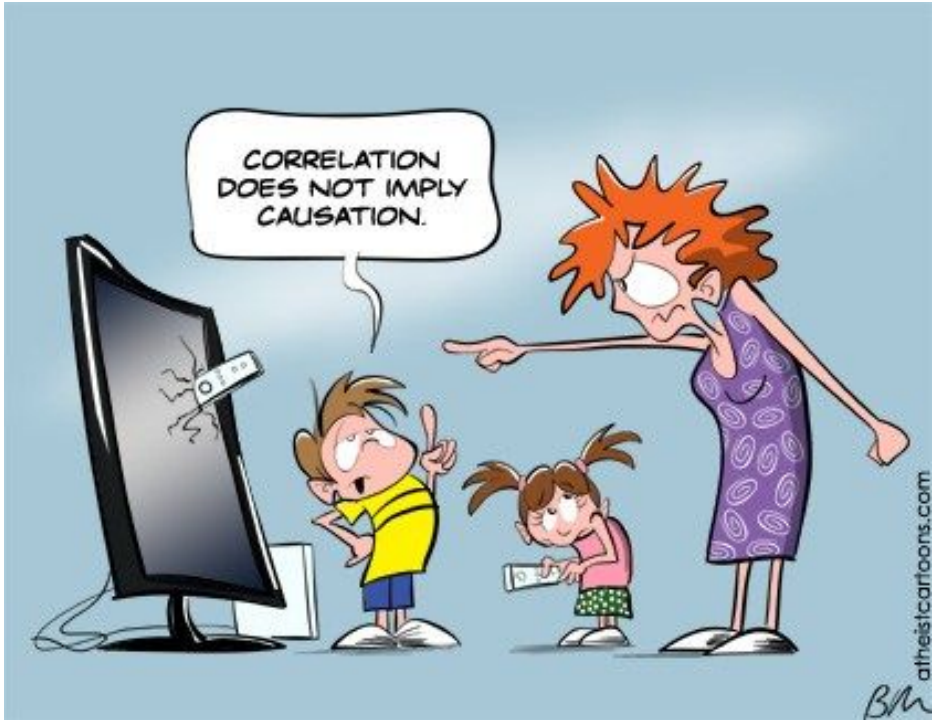
Examples (find the odd one out)

- Economists use Linear Regression to predict the economic growth of a country or state.
- Sports analyst use linear regression to predict the number of runs or goals a player would score in the coming matches based on previous performances.
- Linear Regression can help in analysing whether a political candidate wins an election
- Linear regression analysis can help a builder to predict how much houses it would sell in the coming months and at what price.
- An organisation can use linear regression to figure out how much they would pay to a new joinee based on the years of experience.

Correlation vs Causation

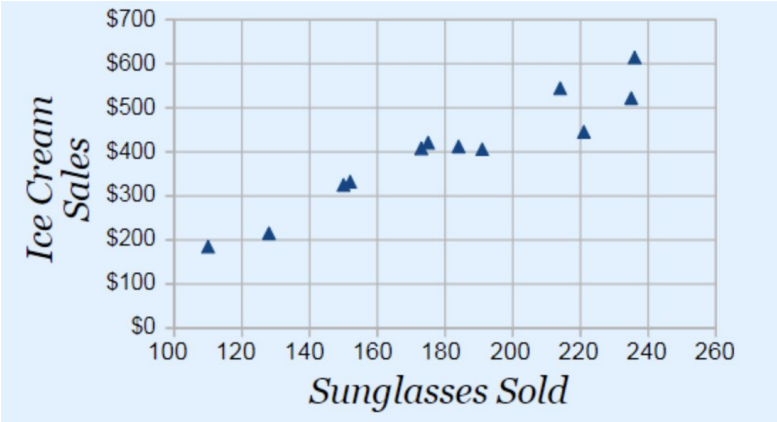
Correlation is a term that refers to the degree of association between two random variables.

Causation is implying that A and B have a cause-and-effect relationship with one another. You're saying A causes B.


CORRELATION DOES NOT IMPLY CAUSATION.
atheistcartoons.com

Let’s take a deeper look

Correlation is a statistical technique which tells us how strongly the pair of variables are linearly related and change together. It does not tell us why and how behind the relationship but it just says the relationship exists.



Let’s take a deeper look

Causation takes a step further than correlation. It says any change in the value of one variable will **cause** a change in the value of another variable, which means one variable makes other to happen. It is also referred as cause and effect.

Example: When a person is exercising then the amount of calories burning goes up every minute. Former is causing latter to happen.

Discussion

Ice cream sales is correlated with homicides in New York (Study).

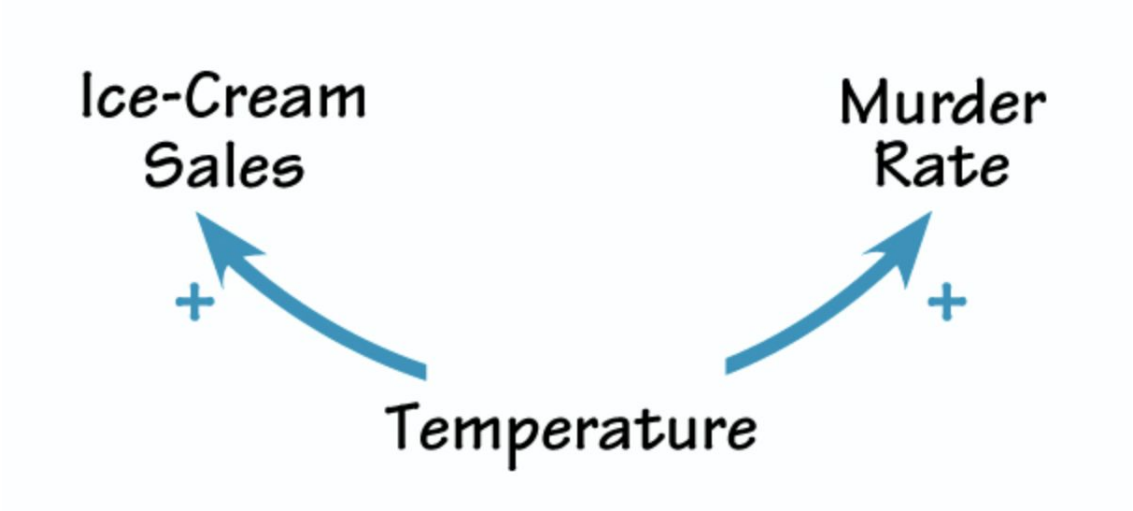
As the sales of ice cream rise and fall, so do the number of homicides. Does the consumption of ice cream causing the death of the people?

Things to consider

Think about the hidden factors that can influence both events in question.

Things to consider

Think about the hidden factors that can influence both events in question.



Single Variable Regression

In a Simple Linear Regression Analysis a straight line approximates the relationship between the dependent variable and the independent variable.

In a Single Variable Regression the relationship is modeled between a single input independent variable

$$Y = aX + b$$

Multi variable Regression

Multi Variable Linear Regression is a model that explores the relationship between multiple independent input variables (feature variables) and an output dependent variable.

$$Y = aX\_1 + cX\_2 + b$$

A few key points

Linear Regression is:

- Fast and easy to model and is particularly useful when the relationship to be modeled is not extremely complex and if you don’t have a lot of data.
- Very intuitive to understand and interpret.
- Linear Regression is very sensitive to outliers.

Regression Coefficients

$$Y = aX + b$$

‘*a*’ - **Regression Coefficient** - tells about the change in the value of dependent variable (Y) corresponding to the unit change in the independent variable (X)

‘*b*’ - is the intercept (the value of ***y*** when ***x*** = 0).

Outliers

An outlier is an observation that lies outside the overall pattern of a distribution

If a value is a certain number of standard deviations away from the mean, that data point is identified as an outlier. The specified number of standard deviations is called the threshold. The default value is usually 3*

*In most cases it will be up to the researcher to define the threshold

Interpreting Results

Y = aX
(No intercept)

	coef	std err
r_all_rs_third	0.9905	0.000

Y = aX+b
(Intercept included)

	coef	std err
const	632.6929	19.936
r_all_rs_third	0.7094	0.009

Interpreting Results

Y = aX
(No intercept)

	coef	std err
r_all_rs_third	0.9905	0.000

Y = aX+b
(Intercept included)

	coef	std err
const	632.6929	19.936
r_all_rs_third	0.7094	0.009

Estimating Results

Things to look for:

RSquare provides a measure of the strength of the linear relationship between the response and the predictor

This statistic, which falls between 0 and 1, measures the proportion of the total variation explained by the model.

Estimating Results

Things to look for:

P-value - tests the null hypothesis* that the coefficient is equal to zero (no effect)

A coefficient that has a low p-value($p<0.05$) is likely to be a meaningful addition to your model

Reference: Hypothesis Testing

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true

Reference: Hypothesis Testing

1. Formulate the Null Hypothesis (Ho) (commonly, that the observations are the result of pure chance) and the alternative Hypothesis (commonly, that the observations show a real effect combined with a component of chance variation).

2. Compute the P-value

3. Compare the P-value to an acceptable significance value (usually 0.1,0.05, 0.01). If P-value < significance value -- the null hypothesis is ruled out, and the alternative hypothesis is valid