Simple Linear Regression

Often we are interested simultaneously in two or more random variables of a random experiment rather than one variable. For example, we might be interested in…

- the voltage signals at two points in a circuit at some specific time
- the repeated measurement of a certain quantity such as the repeated measurement ("sampling") of the amplitude of an audio or video signal that varies with time
- Daily average temperature and power usage
- Number of hours spent studying and test score
- Dosage of a drug and blood pressure

If two random variables are **not independent**, we learned to **quantify the nature and strength of the relationship** between them.

Covariance gives only nature (direction) of a LINEAR relationship.
Correlation gives both strength and nature ——— 〃

There are two types of variables in regression.

Independent variable (x)
- A predictor variable
- A manipulated variable in experiments
- The proposed cause

Response (Dependent) variable (Y)
- An outcome variable (measured at the end of an experiment)
- A non-manipulated variable
- The proposed effect

- Does the power usage depend on the daily average temperature?

  Y = Power usage X = Daily average temperature

- Does the dosage of a drug have an effect on the blood pressure?

  Y = Blood pressure , X = Dosage of drug

- Can the voltage signal at point A in a circuit explain and predict the voltage signal at point B of the same circuit at a given specific time?

  Y = Voltage signal at point B
  X = . ——— 〃 ——— A

"Goals" of Regression

- Describe the relationship between independent (x) and dependent (y) variables
- Based on the identified relationship, predict the y-value for given x-value

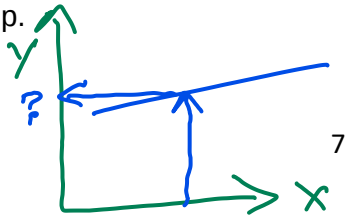
(*) Do NOT use Regression model to predict x-value for a given y-value.

There could be more than one independent variable in a real regression analysis.

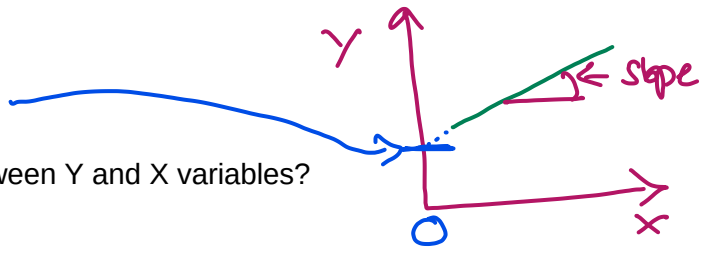
However, only **one** independent variable will be used in the **Simple Regression** analysis.

Further**, linear** regression analysis fits only **a straight line** for the relationship.

"All models are wrong, but some are useful" --- George Box



$Y = mX + b$ (slope, y-intercept)

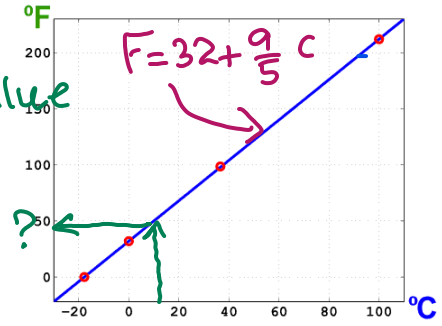


What is the mathematical equation for a straight line between Y and X variables?

$Y = \theta_0 + \theta_1 X$

This is an equation that describes an exact (deterministic) relationship between y and x.

For example, consider the relationship between Temperature in Fahrenheit and temperature in Celsius.



There is exactly one y-value for a single x-value in a deterministic relationship.

There are NO such relationships in Statistics.

In the real life, there are many sources of randomness (or variation):

- Measurement noise
- The linear regression model might be an approximation to a much more complicated and possibly unknown relationship
- Other factors (variables) not used in the regression model

Randomness means the same value of x variable does not always give the same value for the y variable. (Non-deterministic)

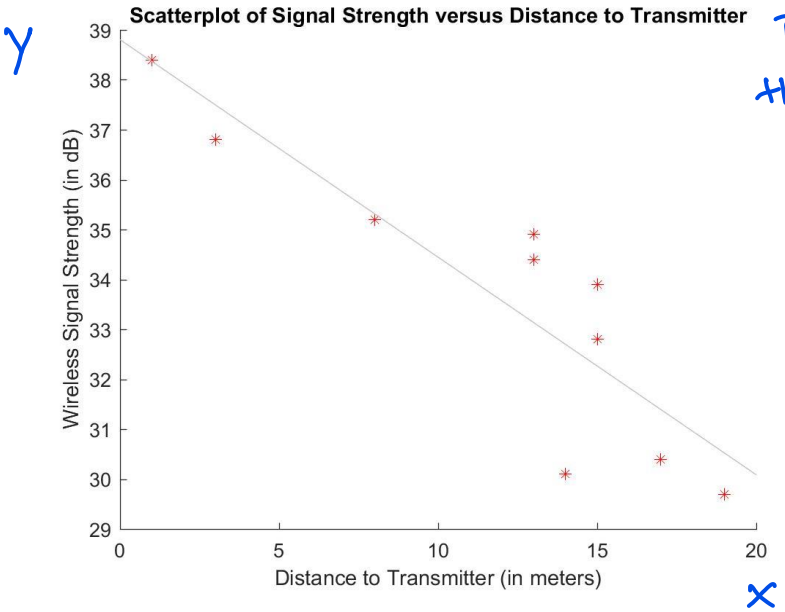
Ex 4: Consider the following 10 measurements on the distance to a transmitter (x) and strength of wireless signal (y).

i	x (m)	y (dB)
1	13	34.4
2	1	38.4
3	17	30.4
4	19	29.7
5	14	30.1
6	15	33.9
7	15	32.8
8	8	35.2
9	13	34.9
10	3	36.8

```
%Create matrix of data %
D = [13 34.4;    1   38.4;   17  30.4;   19  29.7;   14  30.1;
     15  33.9;   15  32.8;   8   35.2;   13  34.9;   3   36.8]

% Create a Scatterplot of Two Numerical Variables %
figure
% scatter(x,y)
scatter(D(:,1), D(:,2), '*r')
xlabel('Distance to Transmitter (in meters)')
ylabel('Wireless Signal Strength (in dB)')
title('Scatterplot of Signal Strength versus Distance to
Transmitter')
lsline % To get the Least Squares line on the scatterplot %

% Compute correlation of Two Numerical Variables %
% corr(x,y)
corr(D(:,1), D(:,2));
fprintf('Correlation Coefficient = %g \n', corr(D(:,1), D(:,2)))
Correlation Coefficient = -0.885064
```



The scatterplot shows that there is a negative medium to strong linear relationship.

This is confirmed by the correlation of -0.885.