early makers

em lyon business school

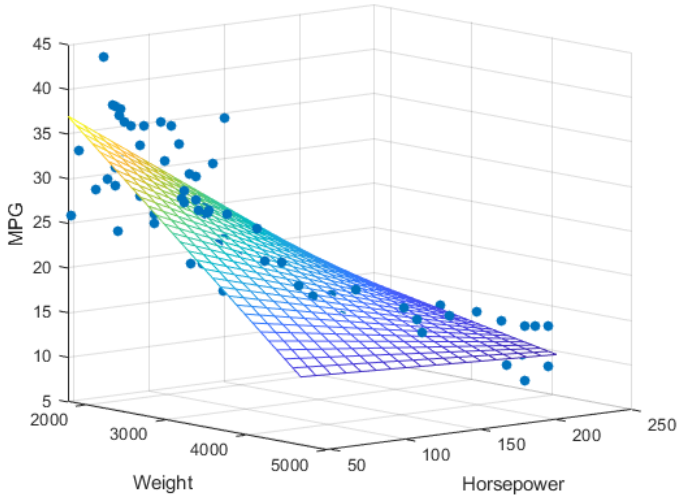
Machine Learning

2. Linear regression

Nicolas Gartner

Part 2: multiple linear regression

Multiple linear regression principle



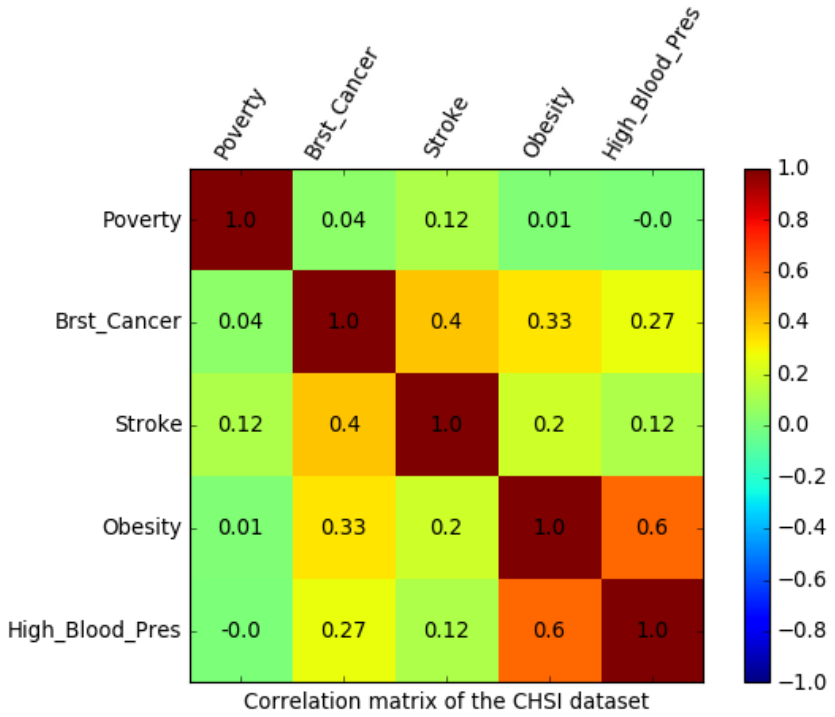
An example of study of relationship between the couple (weight-horsepower) and the consumption of the car in Miles Per Gallon (MPG)

- Similar to simple linear regression
- Uses more than one dimension for the input of the model
- Same metrics can be applied
- Same assumptions are made :
  - There is a linear relationship between the dependent variables and the independent variables.
  - The independent variables are not too highly correlated with each other.
  - Observations are selected independently and randomly from the population.
  - Residuals should be normally distributed with a mean of 0 and variance σ.

Correlation matrix

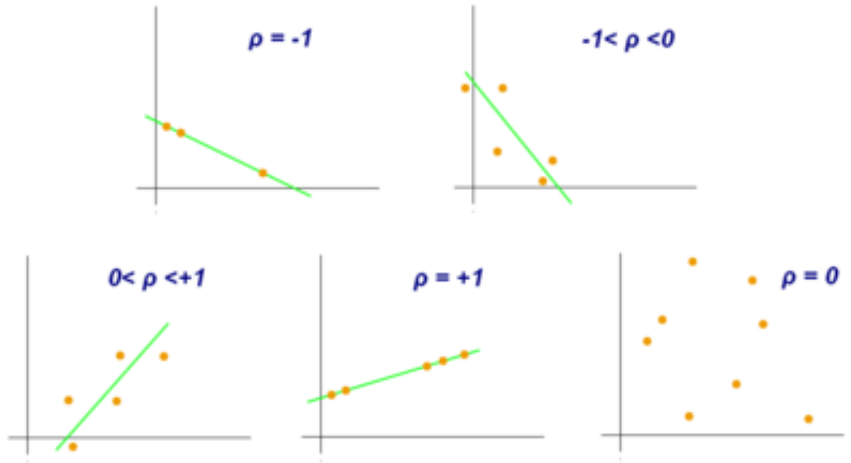
- Table showing correlation coefficients between variables
- Each cell in the table shows the correlation between two variables.
- Very useful to pick up relationship in a dataset.
- Matrixes are diagonally symmetric
- Maximum value is always 1, which can be seen on the diagonal
- Methods :
  - Pearson's Product-Moment Correlation (Most common)
  - Spearman's rank Correlation
  - Kendall's Tau

The 3 methods are available with pandas



Pearson correlation coefficient

- Assesses how well the relationship between two variables can be described using a linear function.
- The correlation coefficient ranges from −1 to 1 :
  - 1: $X = \alpha\, Y + \beta$. All data points lie on a line for which *Y* increases as *X* increases.
  - −1: $X = -\alpha\, Y + \beta$. All data points lie on a line for which *Y* decreases as *X* increases
  - A value of 0 implies that there is no linear correlation between the variables.



- Generic formula:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

n the sample size
x_i and y_i are the individual sample points indexed with i
$\bar{x}$ and $\bar{y}$ are the mean of x and y samples

- Simplified (for the computer) formula

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$ corrected standard deviation

Pearson correlation coefficient

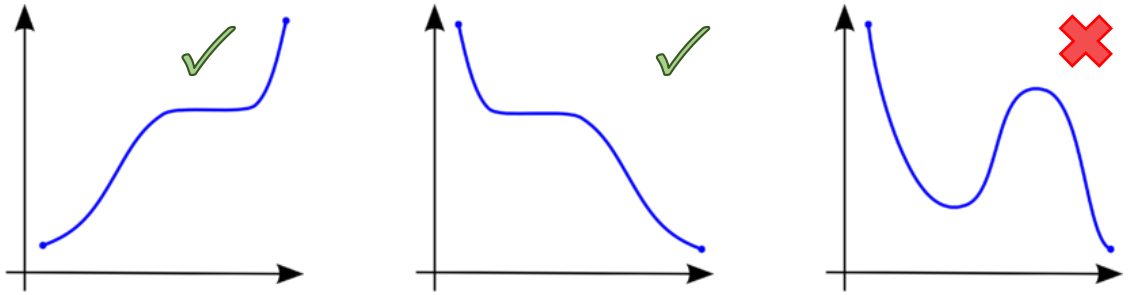
- Additional properties:
  - **Pure number:** It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
  - **Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

- Degree of correlation:
  - **Perfect:** If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
  - **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
  - **Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
  - **Low degree:** When the value lies below ± .29, then it is said to be a small correlation.
  - **No correlation:** When the value is zero.

Spearman correlation coefficient

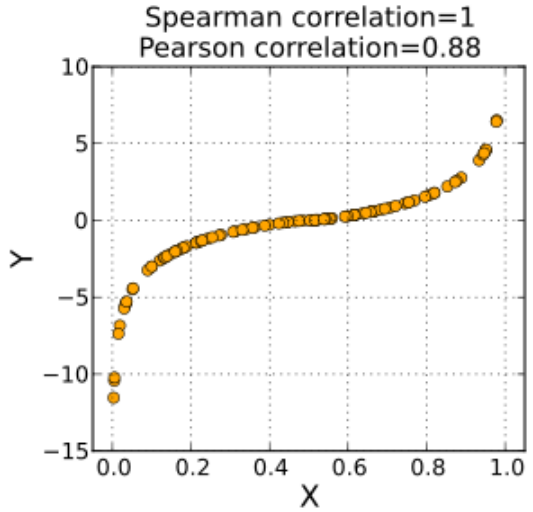
- Assesses how well the relationship between two variables can be described using a monotonic function.
  - A monotonic function is a function that is entirely increasing or decreasing

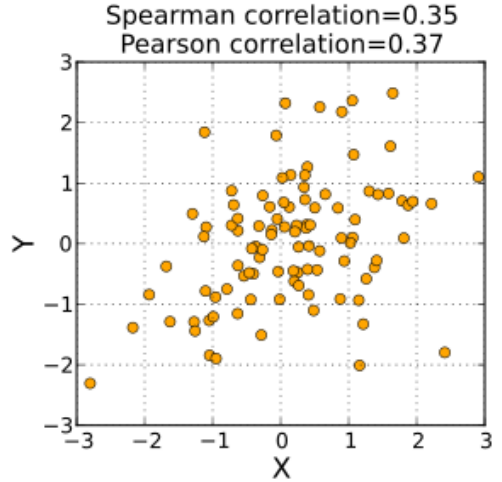


- Evolves between +1 and -1 just as Pearson coefficient
- Formula:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

With $d_i = rk(x_i) - rk(y_i)$ the difference between the two ranks of the observation



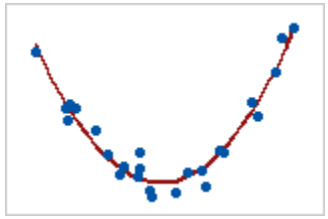


Spearman correlation coefficient

- An example to get what the rank is:

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

- Spearman and Pearson coefficients might be used combinedly.
- If both coefficients are 0, then it doesn't mean that there is no relationship



- In this example both coefficient values are 0.
- This might mean that the relationship is non-linear

Kendall tau-a coefficient

- Is like Spearman coefficient a way to assess the monotony of the curve
- Evolves between -1 and 1
- Works with pairs of data/observations

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
…	…
x_n	y_n

(x_i, y_i) (x_j, y_j) With $i < j$

Concordant if:

$x_i > x_j \; and \; y_i > y_j$ | $x_i < x_j \; and \; y_i < y_j$

- Formula:

$$\tau = 2\frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)} = \frac{n_c - n_d}{\binom{n}{2}}$$

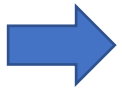
Binomial coefficient

- There are 3 derived coefficients (Tau-a, Tau-b and Tau-c). Tau-a is the “classic” version shown above, Tau-b tries to consider ties and Tau-c considers ties and differences in that amount of data between X and Y.

An example of how it works

- List of rankings between two interviewers and candidates

Candidate	Interviewer 1	Interviewer 2
A	1	1
B	2	2
C	3	4
D	4	3
E	5	6
F	6	5
G	7	8
H	8	7
I	9	10
J	10	9
K	11	12
L	12	11



Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	0
B	2	2	10	0
C	3	4	8	1
D	4	3	8	0
E	5	6	6	1
F	6	5	6	0
G	7	8	4	1
H	8	7	4	0
I	9	10	2	1
J	10	9	2	0
K	11	12	0	1
L	12	11		
Totals:			61	5

$$\tau = \frac{n_c - n_d}{\frac{n*(n-1)}{2}} = \frac{61-5}{\frac{12*11}{2}} = 0.85$$

Video

True Artificial Intelligence will change everything | Juergen Schmidhuber

https://youtu.be/-Y7PLaxXUrs