# Régression Linéaire

et Python

# COURS PRÉCÉDENT

- Révisions de python
- Différence entre Data Science, Machine Learning et analyse prédictive
- Approche statistique vs approche machine learning
- Déroulement d'un projet de Data science
- Supervisée vs non-supervisée
- Régression vs Classification
- Anaconda, Python et Jupyter

# RÉGRESSION LINÉAIRE

- Régression linéaire
    - OLS, Moindres carrés
    - Modélisation
    - Univariable & multivariables
- Interprétation des résultats
    - Mean Square Error (MSE)
    - P-value, Interval de confiance, $R^2$, $R^2_{adj}$
    - Confonders et multi-collinearité
- Hypothèses et vérification
    - Linéarité: Définition et tests
- Statsmodel

### PROJET KAGGLE

### PYTHON

# LAB

Lab pandas et exploratio
arbres.csv

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

# House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,200 teams · Ongoing

Overview Data Kernels Discussion Leaderboard Rules **Join Competition**

Overview

**Description**

**Evaluation**

**Tutorials**

**Frequently Asked Questions**

## Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

## Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

### CLASSIFICATION: QUANTITATIF

La variable à prédire est **discrète**

#### CAS BINAIRE

- Achat, resiliation, click
- Survie, maladie, succes examen, admission,
- Positif ou negatif
- Spam, fraude

#### MULTI CLASS - MULTINOMIALE

- Catégories, types (A,B,C),
- Positif, neutre ou negatif
- Espèces de plantes d'animaux, ...
- Pays, planetes

#### ORDINALE

- Notes, satisfaction, ranking

### REGRESSION: QUALIT

La variable à prédire est

- Age, taille, poids,
- nombre d'appels, d
consommation
- Température, Salai
- Probabilité d'une a
- Temps, délai, retar

## TAILLE EN FONCTION DE L'AGE DES ENFANTS

On mesure la taille des enfants dans une ecole et leur age. La taille croit avec l'age. On peut écrire

$$\text{Taille} = f(\text{Age})$$

## REGRESSION UNIVARIABLE

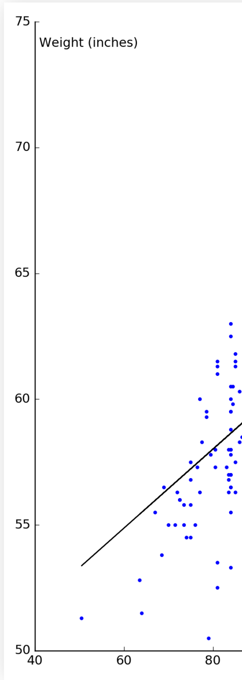
On modélise cette fonction par une relation linéaire de la forme:

$$\hat{\text{Taille}} = a * \text{Age} + b$$

où $\hat{\text{Taille}}$ est la taille estimée.

On cherche à connaitre les paramètres $(a, b)$ qui donnent la meilleure approximation de la réalité entre la taille et l'age.

Pour trouver ces paramètres on utilise une méthode dite des **moindres carrés** ou **Ordinary Least-Squares (OLS)**.



# REGRESSION LINÉAIRE

Nous avons $n$ échantillons:

- Une variable prédictrice $x = [x_1, \ldots, x_n]$
- Et une variable cible $y = [y_1, \ldots, y_n]$

On veut trouver les *meilleurs* $a$ et $b$ pour lesquels

$$\hat{y}_i = a * x_i + b$$

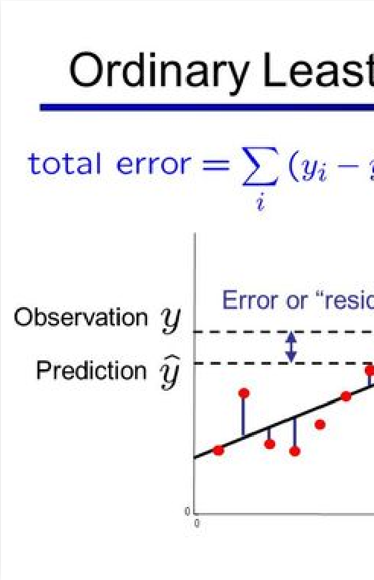
l'erreur de prédiction $e_i$ soit minimale:

$$e_i = |y_i - \hat{y}_i| = |y_i - (a * x_i + b)|$$

Les résidus $e_i$ représente
vrais valeurs $y_i$ et leur es
réduire cette distance.

Pour cela on chercher à
carrés des résidus (aussi

$$||y - \hat{y}||^2 = \sum_{i=0}^{n}$$



# PETIT RAPPEL DES NORMES

## NORME QUADRATIQUE $\mathbf{L^2}$

$$||x|| = \sqrt{x_1^2 + .... + x_n^2}$$

## NORME $\mathbf{L^1}$ OU NORME EN VALEUR ABSOLUE

$$|x| = |x_1| + .... + |x_n|$$

## NORME INFINIE $\mathbf{L^\infty}$

$$|x|_\infty = max[|x_1|, \dots, |x_n|]$$

# FONCTION DE COUT

On a ce qu'on appelle une **fonction de cout** $L(a, b)$:

$$L(a,b) = ||y - \hat{y}||^2 = \sum_{i=0}^{n} [y_i - (a * x_i + b)]^2$$

C'est fonction quadratique donc convexe.

Par conséquent pour trouver son minima, il faut trouver les valeurs de $a$ et $b$ qui annule la dérivée $0$ .

Cela donne 2 équations à 2 inc
exacte est:

$$\hat{\beta} = (x^T . x$$

avec

- $\hat{\beta} = \{a, b\}^T$
- $x = [x_1, \ldots, x_n]$
- $y = [y_1, \ldots, y_n]$

# REGRESSION MULTINOMIAL

### PLUSIEURS PREDICTEURS

On a maintenant $m$ predicteurs et toujours $n$ échantillons.

Pour chaque échantillon, on a la modélisation suivante:

$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

ou plus simplement

$$\hat{y} = \beta X$$

avec

- $X = [(x_{i,j})]$ est une matrice de taille $n$ par $m$
- $y = [y_1, \ldots, y_n]$ vecteur de $n$ échantillons

On veut trouver les n+1 coeffici

$$\beta = [\beta_0, \beta_1,$$

qui minimisent la fonction de c

$$L(\beta) = ||y$$

La solution de cette équation e

$$\hat{\beta} = (X^T . X$$

# REGRESSION LINÉAIRE

A) N samples avec M variables:

$$y_i = \sum_k \beta_k * X[i,k] + \sigma^2$$

$$y = \beta * X + \sigma^2$$

B) Regression weights:

$$\hat{\beta} = (X^T.X)^{-1}X^T y$$

C) Prédiction

$$\hat{y}_i = \sum_k \beta_k * X[i,k] + \sigma^2$$

# PYTHON

A)

```
X, y = make_regression(n_samples=N
```

B)

```
beta_hat = np.linalg.inv(X.T.dot(X
```

C)

```
yhat = X[:, 0]* beta[0] + X[:,1] *
```

- ou si $M > 2$:

```
yhat = [0 for i in range(N)]

for k in range(M):
    yhat += X[:, k]* beta[k]
```

# NOTEBOOK - DEMO

02 Linear Regression Exact.ipynb

# METRIQUES DE SCORING

## ERREUR ABSOLU (MAE) (L1)

Valeur absolue de la difference entre la prédiction et les vraies valeurs

$$MAE = \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

```
e = np.mean( np.abs(y - yhat) )
```

## ERREUR QUADRATIQUE (MSE) (L2)

$$MSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

```
```e = np.mean( (y - yhat)**2 )
```
```
```
```

# RÉGRESSION LINÉAIRE AVEC STATSMODEL

On va estimer les coefficients non plus directement mais avec la méthode OLS.

On aura plus d'information sur les coefficients de régression:

- leur importance relative
- leur fiabilité
- leur impact quantitatif

On utilise la librairie

- Statsmodel librairie Python pour une approche statistique de l'analyse de données.
- Intégrée avec pandas et numpy

Sur un vrai dataset: **Mileage pe**
**various cars** disponible sur
https://www.kaggle.com/ucim

A prédire:

- mpg: continuous

Les variables

- cylinders: multi-valued dis
- displacement: continuous
- horsepower: continuous
- weight: continuous
- acceleration: continuous

On ne prends pas en compte:

- model year: multi-valued
- origin: multi-valued discre
- car name: string (unique f

# STATSMODEL

```
In [40]: smf.
          GEE            GLS            Logit          MNLogit            nominal_gee    ordinal_gee    Poisson        QuantRe
          gee            gls            logit          mnlogit            NominalGEE     OrdinalGEE     poisson        quantre
          GLM            GLSAR          MixedLM        NegativeBinomial   OLS            PHReg          Probit         RLM
          glm            glsar          mixedlm        negativebinomial   ols            phreg          probit         rlm
```

# NOTEBOOK PYTHON

```
import pandas as pd
import statsmodels.formula.api as smf

df = pd.read_csv('../data/autos_mpg.csv')
lm = smf.ols(formula='mpg ~ cylinders + displacement + horsepower + weight + acceleration + origin ', data=df).fit()
lm.summary()
```

| OLS Regression Results | |
|---|---|
| **Dep. Variable:** | mpg |
| **Model:** | OLS |
| **Method:** | Least Squares |
| **Date:** | Sat, 22 Sep 2018 |
| **Time:** | 17:58:03 |
| **No. Observations:** | 398 |
| **Df Residuals:** | 391 |
| **Df Model:** | 6 |
| **Covariance Type:** | nonrobust |

# RÉSULTATS

- **Dep. Variable**: La variable
- **Model**: Le modèle
- **Method**: La méthode utilis
- **No. Observations**: Le nom échantillons
- **DF Residuals**: Degré de lib d'échantillons - nombre d
- **DF Model**: Nombre de pré

## GOODNESS OF

| | |
|---|---|
| **R-squared:** | 0.717 |
| **Adj. R-squared:** | 0.713 |
| **F-statistic:** | 165.5 |
| **Prob (F-statistic):** | 4.84e-104 |
| **Log-Likelihood:** | -1131.1 |
| **AIC:** | 2276. |
| **BIC:** | 2304. |

- **R-squared**: The coefficien
statistical measure of how
approximates the real dat
- **Adj. R-squared**: The above
the number of observation
freedom of the residuals
- F-statistic: A measure how
mean squared error of the
squared error of the residu
- Prob (F-statistic): The prob
the above statistic, given t
are unrelated
- Log-likelihood: The log of
- AIC: The Akaike Informatio
likelihood based on the nu
the complexity of the mod
- BIC: The Bayesian Informa
the AIC, but has a higher p
more parameters.

# $\mathbf{R^2}$

Soit la moyenne de la variable cible :

$$\bar{y} = \frac{1}{n} \sum_i y_i$$

et la somme des carrés de la variable cible centrée :

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

La somme des carrés des résidus :

$$SS_{\text{res}} = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

# DÉFINITION

$R^2$ est la proportion des variati
est prédite grace aux prédicteu

On définit $R^2$ par

$$R^2 = 1 -$$

On a

$$0 < R^2$$

# R2 DOES NOT INDICATE WHETHER:

- the independent variables are a cause of the changes in the dependent variable;
- omitted-variable bias exists;
- the correct regression was used;
- the most appropriate set of independent variables has been chosen;
- there is collinearity present in the data on the explanatory variables;
- the model might be improved by using transformed versions of the existing set of independent var
- there are enough data points to make a solid conclusion.

et surtout

- plus on ajoute de variable plus $R^2$ augmente meme quand les variables ne sont pas vraiment signi

# $\mathbf{R^2_{adj}}$

On ajuste pour prendre en compte la complexité du modele:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

avec

- $p$ le nombre de prédicteurs
- $n$ le nombre d'échantillons

En accroissant le nombre de pr
souvent R^2.

Mais $R^2_{adj}$ compense la comple

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 42.7111 | 2.693 | 15.861 | 0.000 | 37.417 | 48.005 |
| **cylinders** | -0.5256 | 0.404 | -1.302 | 0.194 | -1.320 | 0.268 |
| **displacement** | 0.0106 | 0.009 | 1.133 | 0.258 | -0.008 | 0.029 |
| **horsepower** | -0.0529 | 0.016 | -3.277 | 0.001 | -0.085 | -0.021 |
| **weight** | -0.0051 | 0.001 | -6.441 | 0.000 | -0.007 | -0.004 |
| **acceleration** | 0.0043 | 0.120 | 0.036 | 0.972 | -0.232 | 0.241 |
| **origin** | 1.4269 | 0.345 | 4.136 | 0.000 | 0.749 | 2.105 |

# COEFFICIENTS

La deuxième partie des résulta
et leur fiabilité.

- **coef**: La valeur estimée de
- **P > |t|**: la probabilité que l
alors qu'en fait le coefficie
- **[95.0% Conf. Interval]**: l'in
l'estimation du coefficient
- **std err**: l'erreur d'estimati
- **t-statistic**: une mesure de
statistique de chaque coef

# P-VALUE

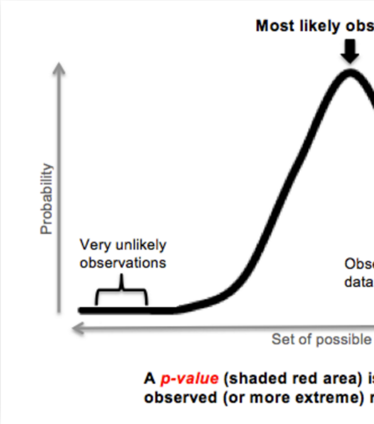
On a 2 hypothèses:

1. [NULL] ce que l'on observe est du au hasard
2. [ALT] ce que l'on observe n'est pas du au hasard (il y a une relation)

La p-value est la probabilité que ce que l'on observe est du au hasard.

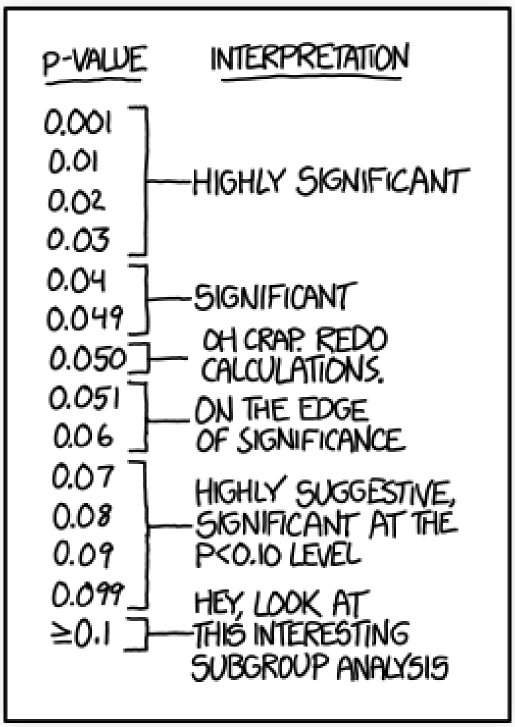
Si la p-value est faible, on rejete l'hypothèse NULL.

Ce qui ne veut pas dire que la valeur du coefficient est la bonne. (ca serait trop simple) mais simplement que il y a bien une relation entre le predicteur et la variable cible.

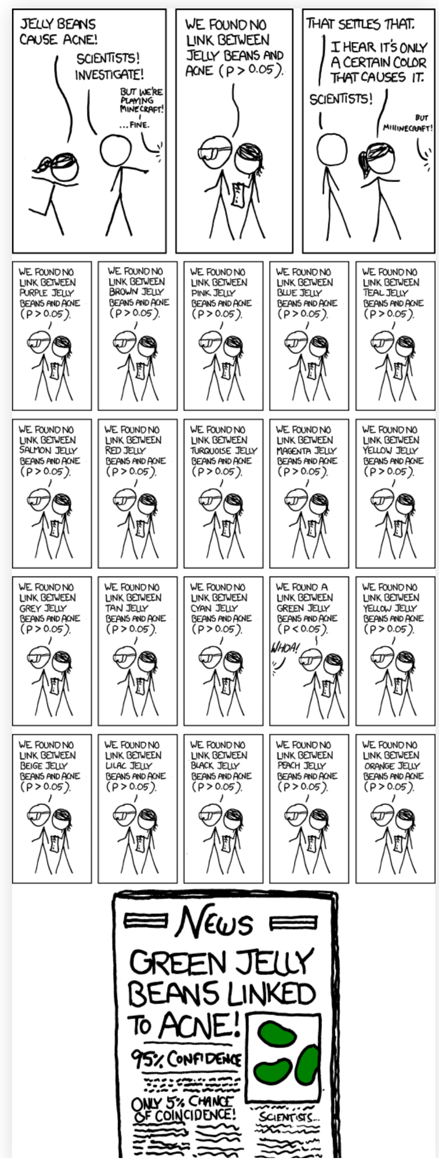



The p-value represents t
coefficient is actually zer

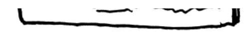
- Si $P_{value} > 0.05$ alo
  chance que l'hypot
  **ne peut pas la rejet**
- si $P_{value} < 0.05$ a l
  chance pour que l'h
  => **on peut la rejete**




P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...



# MULTINOMIALE

Que se passe t il quand on filtre certains predicteurs ?

# CONDITIONS SUR LES DONNÉES

Pour qu'une régression linéaire soit possible et fiable, il faut que les données vérifient les conditions sui

- **Linearite**: la relation entre les predicteurs et la cible est lineaire
    - On peut tester avec des scatter plots
- **Normality**: Les variables ont une distribution normale
    - test: QQ plot
    - ou Kolmogorov-Smirnov test
    - correction: log ou box-cox
- **Independence**: no or little multicollinearity between variables
    - test: Correlation matrix
- **Homoscedasticity:** for a given variable the low and high range have the same statistical properties residuals
    - test: Chunk data and Check Variance
- All **Confounders** accounted for

http://www.statisticssolutions.com/assumptions-of-linear-regression/

https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutic

https://www.kaggle.com/questions-and-answers/57571 What does Linear Relationship means? :- You ca
scatter plots, if there is little to no linearity in the scatter plot between your dependent and Independen
assumption doesn't hold. Linearity regression assumption requires all variables to be normal, how to ch
assumption with a Q-Q plot, if your data deviates substantially from the line on the Q-Q plot, then this a
check for little to no multicollinearity, why is multicollinearity a problem? :- Multicollinearity generally o
correlations between two or more predictor variables. A principal danger of such data redundancy is th
regression analysis models. The best regression models are those in which the predictor variables each
dependent (outcome) variable but correlate at most only minimally with each other. Such a model is of
will be statistically robust (that is, it will predict reliably across numerous samples of variable sets draw
population).

# Correlation

# RAPPEL PEARSON COEFFICIENT

Etudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques numériques, c'est étud
qui peut exister entre ces variables.

Il y a différentes façon de calculer la corrélation de 2 variables.

La plus commune est Pearson Correlation

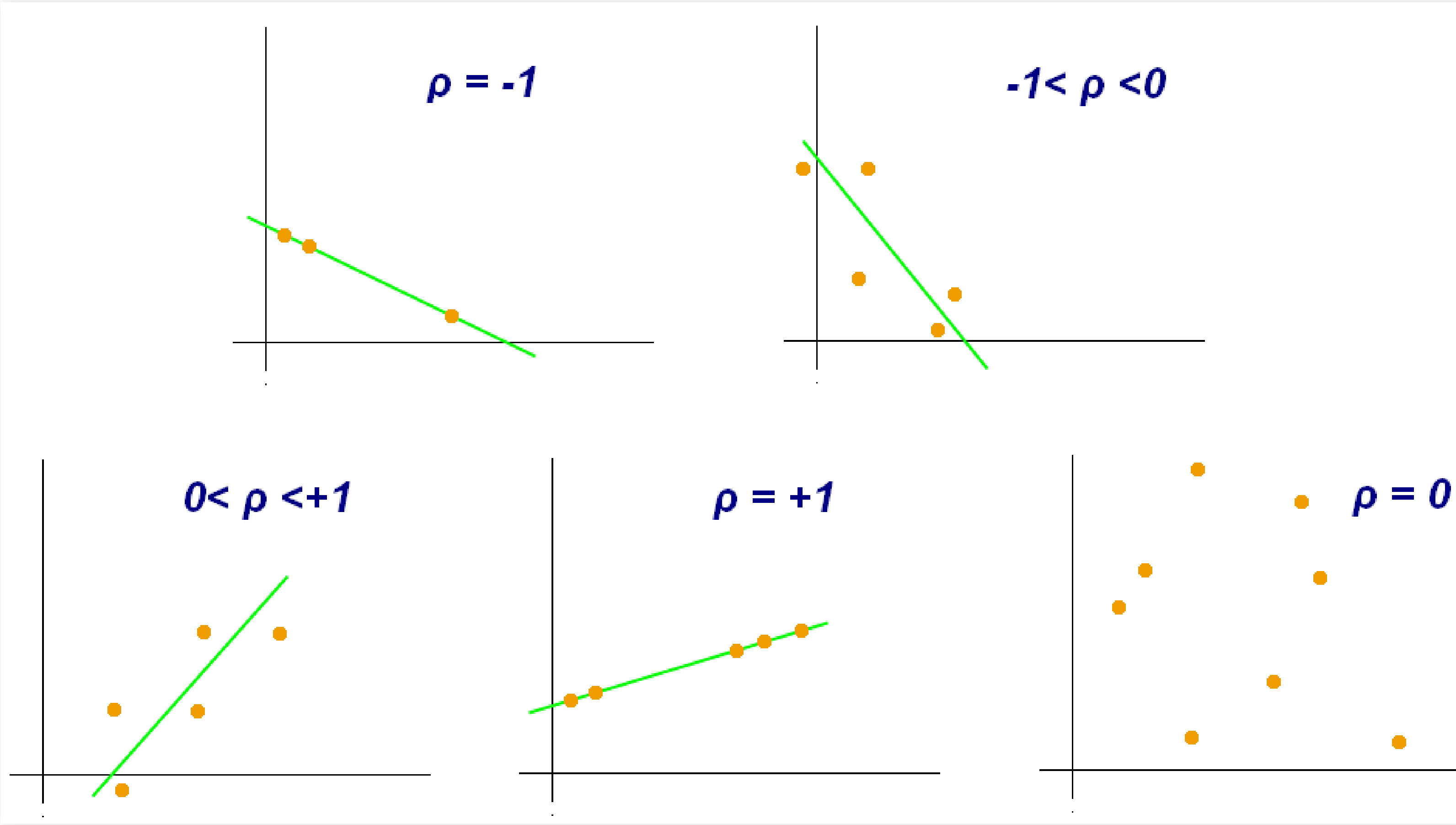
Qui se calcule suivant :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

où :

- $n$ nombre d'échantillons
- $x_i, y_i$ les échantillons
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la moyenne; de meme pour $\bar{y}$

# CORRELATION



# CORRELATION

On va regarder l'influence de la correlation entre les predicteurs

```
df.corr()
```

Les prédicteurs `horsepower` et `weight` sont très corrélés, `displacement` et `cylinders` aussi.

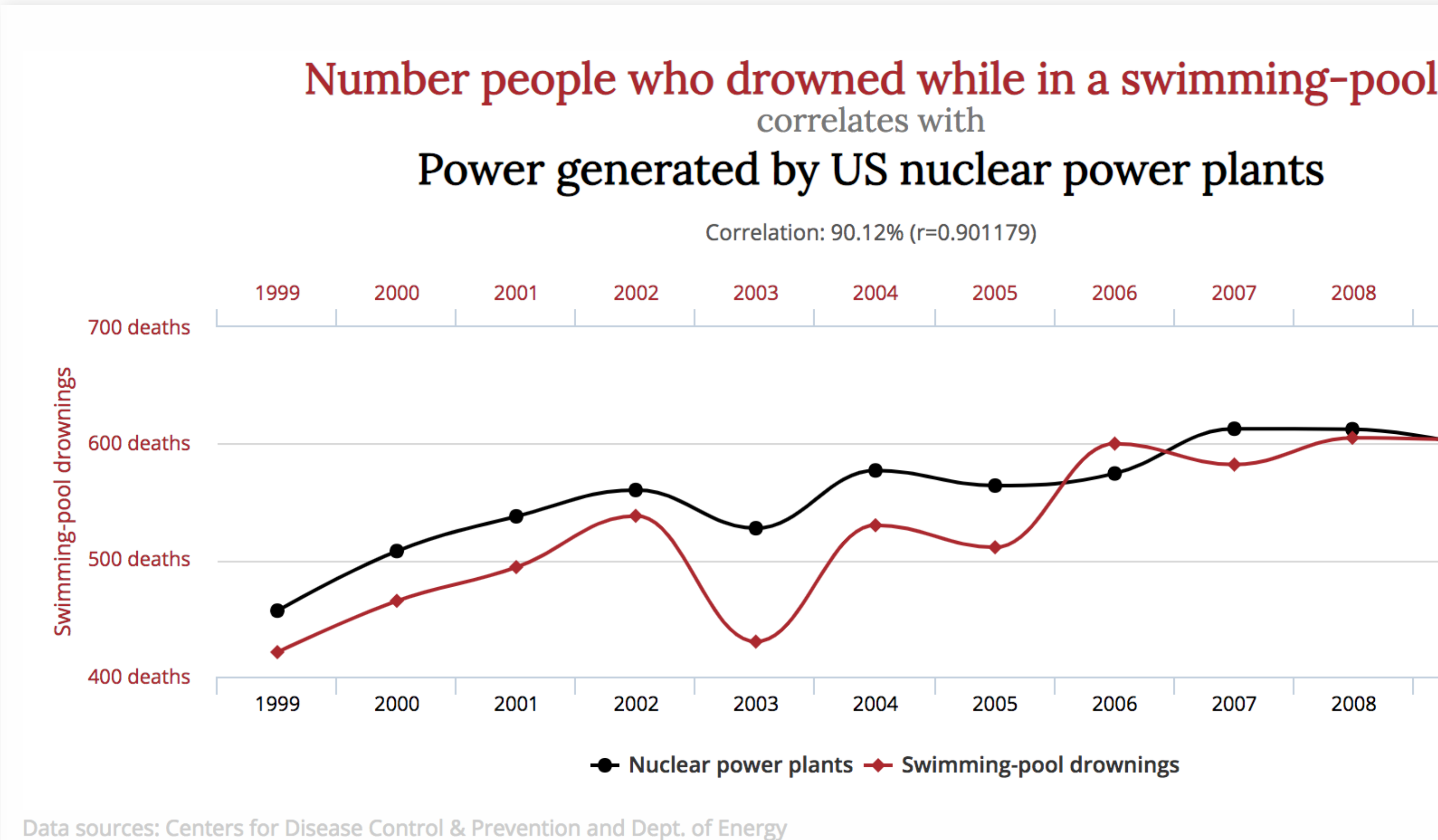
```
In [4]: df.corr()
```

Out[4]:

| | Unnamed: 0 | mpg | cylinders | displacement | horsepower | weight | acceleration |
|---|---|---|---|---|---|---|---|
| **Unnamed: 0** | 1.000000 | 0.585131 | -0.363040 | -0.386976 | -0.417861 | -0.318869 | 0.287634 |
| **mpg** | 0.585131 | 1.000000 | -0.775396 | -0.804203 | -0.771437 | -0.831741 | 0.420289 |
| **cylinders** | -0.363040 | -0.775396 | 1.000000 | 0.950721 | 0.838939 | 0.896017 | -0.505419 |
| **displacement** | -0.386976 | -0.804203 | 0.950721 | 1.000000 | 0.893646 | 0.932824 | -0.543684 |
| **horsepower** | -0.417861 | -0.771437 | 0.838939 | 0.893646 | 1.000000 | 0.860574 | -0.684259 |
| **weight** | -0.318869 | -0.831741 | 0.896017 | 0.932824 | 0.860574 | 1.000000 | -0.417457 |
| **acceleration** | 0.287634 | 0.420289 | -0.505419 | -0.543684 | -0.684259 | -0.417457 | 1.000000 |
| **model year** | 0.996800 | 0.579267 | -0.348746 | -0.370164 | -0.411651 | -0.306564 | 0.288137 |
| **origin** | 0.199702 | 0.563450 | -0.562543 | -0.609409 | -0.453669 | -0.581024 | 0.205873 |

# CORRELATION $\neq$ CAUSALITÉ

http://www.tylervigen.com/spurious-correlations



Number people who drowned while in a swimming-pool
correlates with
Power generated by US nuclear power plants
Correlation: 90.12% (r=0.901179)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
700 deaths
600 deaths
500 deaths
400 deaths
Swimming-pool drownings
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
Nuclear power plants
Swimming-pool drownings
Data sources: Centers for Disease Control & Prevention and Dept. of Energy

# REGRESSION AND CAUSATION:

For regression coefficients to have a causal interpretation we need both that

- the linear regression assumptions hold: linearity, normality, independence, homoskedasticity
- and that all confounders of, e.g., the relationship between treatment A and Y be in the model.

Not the same thing

Calculating Correlation: easy

Demonstrating and Quantifying Causation: Causal Inference: Not so easy

=> However most common strategy is to find not causality but correlation through linear regression whi
causality under strong assumptions on the covariates.

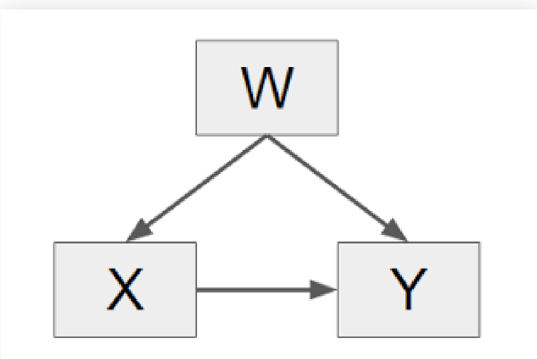
**Works under VERY strong assumptions**

# CONFONDERS

facteurs potentiels de confusion

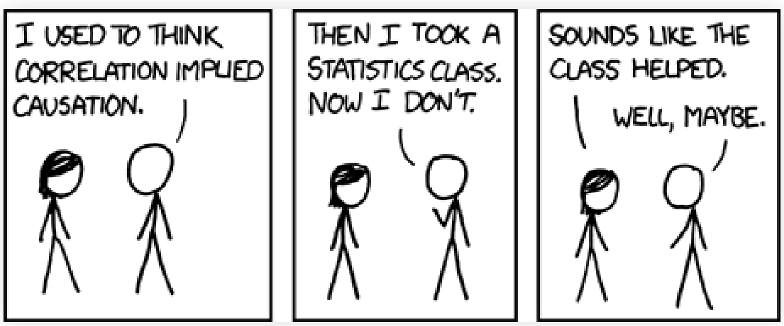
https://www.r-bloggers.com/how-to-create-confounders-with-regression-a-lesson-from-causal-inferen

http://www.statisticshowto.com/experimental-design/confounding-variable/




- Relationship between **ice-cream consumption** and number of **drowning deaths** for a given period

**Confounding: ?**


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

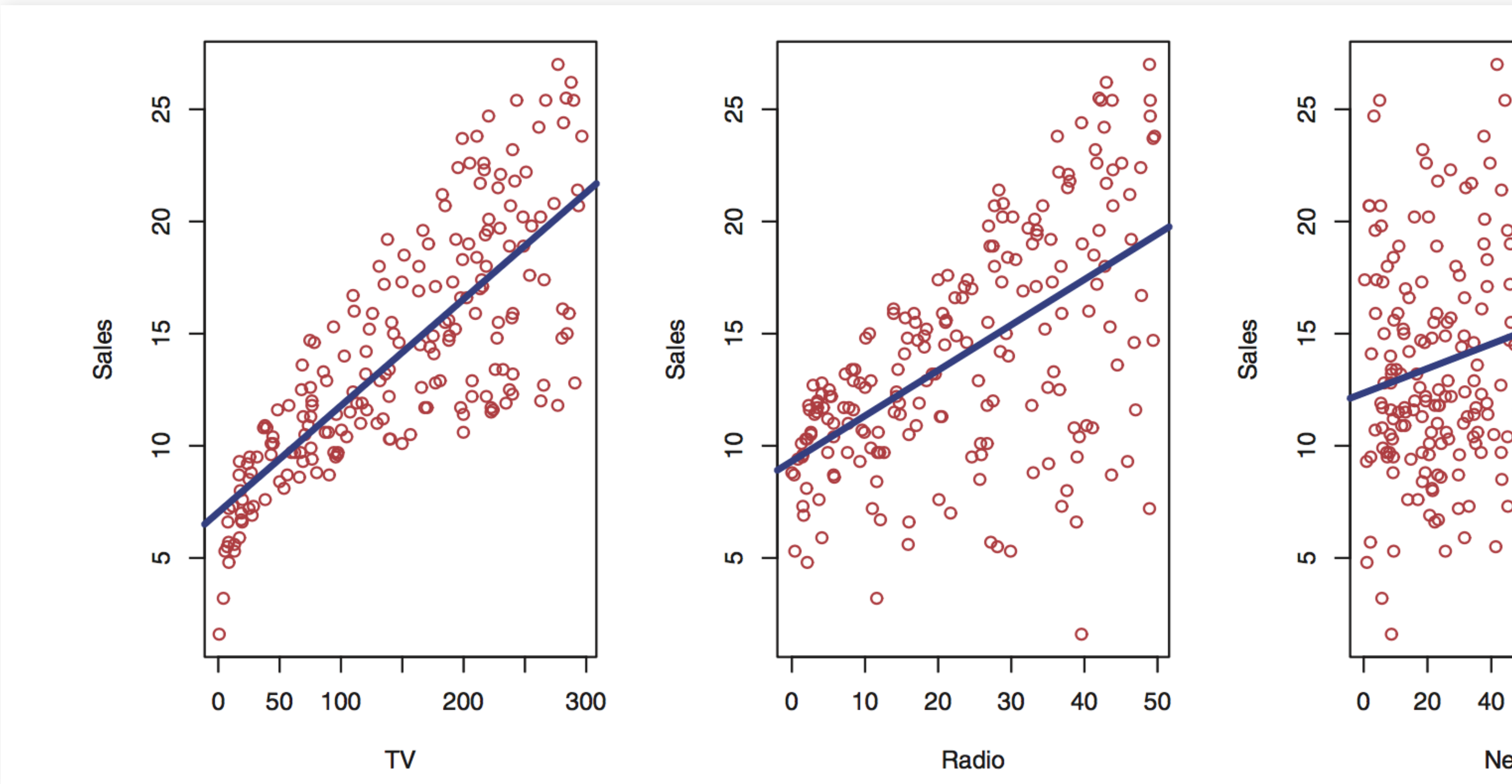
# Récapitulatif

# RÉCAPITULATIF

- Regression lineaire, simple et explicite
- Attention à ce que les predicteurs soient decorrélés
- R^2 ajusté au lieu de R^2

# LAB DE CETTE APRES MIDI

Regression lineaire sur le dataset *advertising*



# QUESTIONS

# LIENS ET RESOURCES

- Régression linéaire en python sur le site de Xavier Dupré
- OLS sur wikipedia: tres complet