data science
for (physical) scientists V

the ABC of regression

dr.federica bianco | fbb.space | fedhere | fedhere

recap 1

Uncertainties

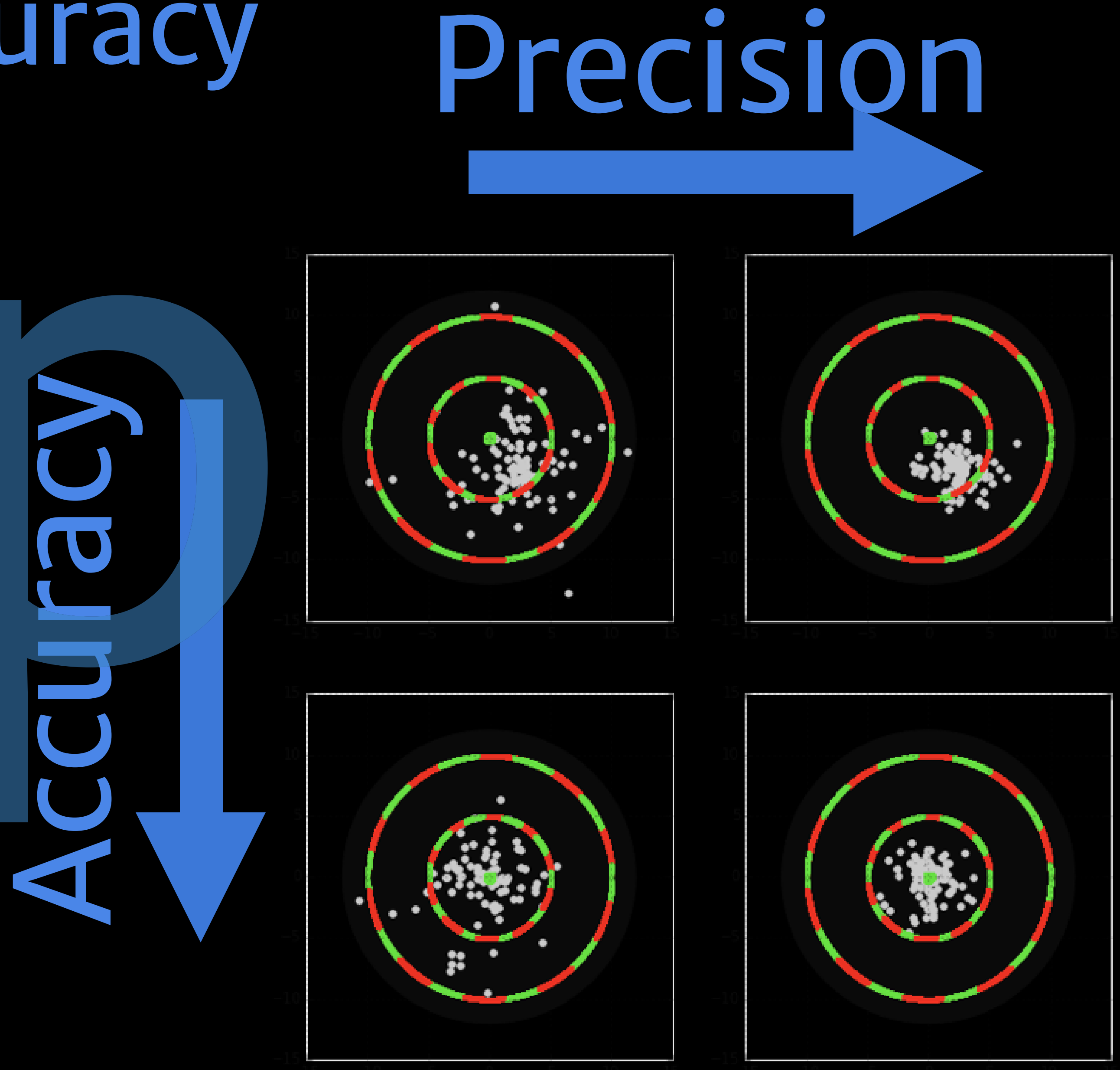
- *types of uncertaintes: Statistical and Systematics*

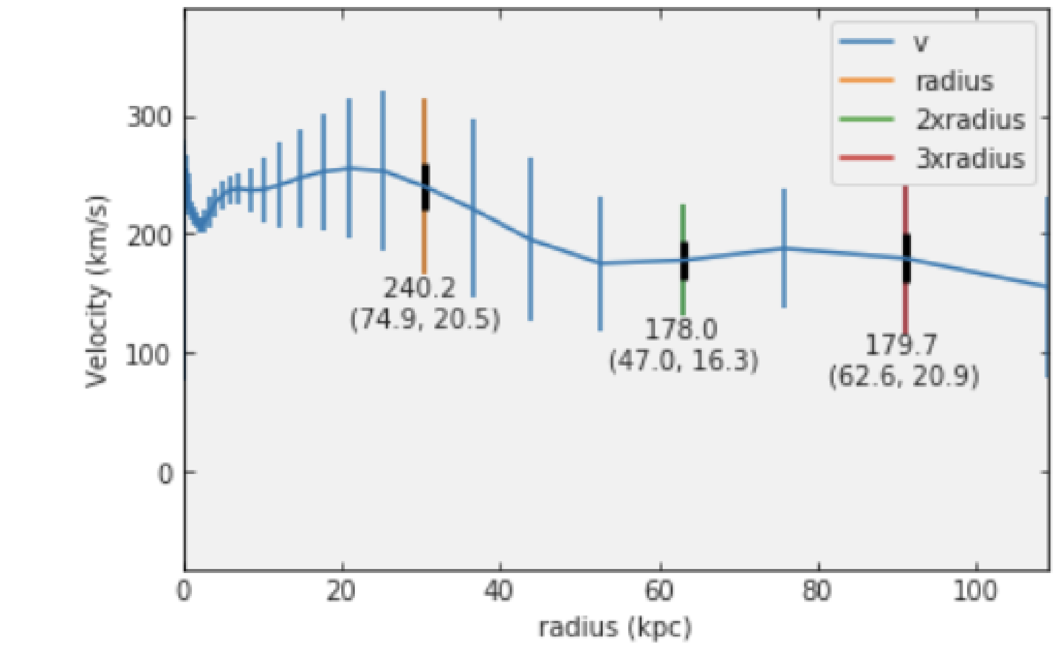
Stochastic vs Systematics

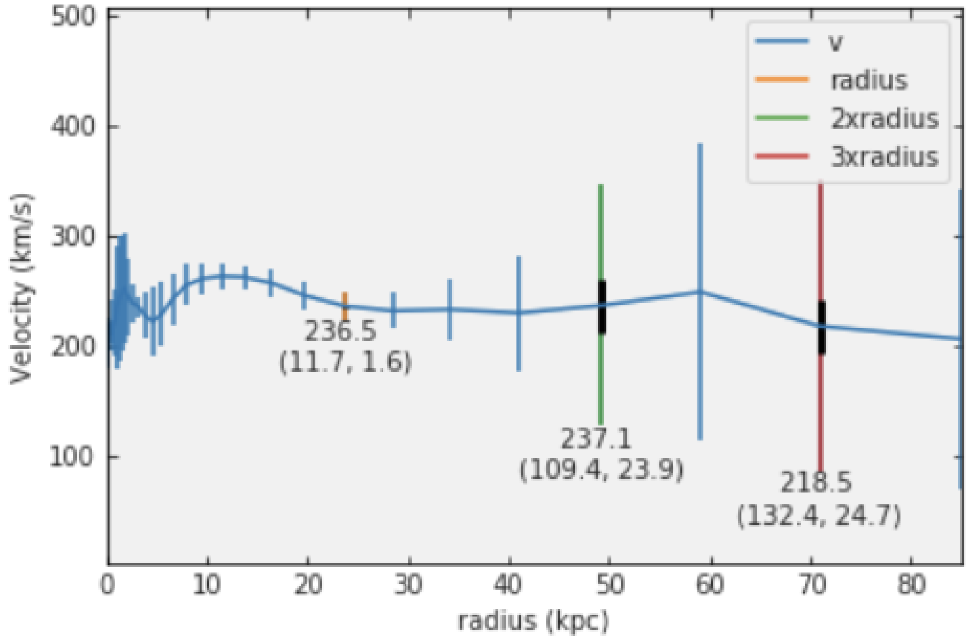
Systematic	Statistical
Biases the measurement *in one direction*	No preferred direction
Affects the sample regardless of the size	Shrinks with the sample size (typically as N)
Any distribution (usually we use Gaussian though)	Typically Gaussian or Poisson

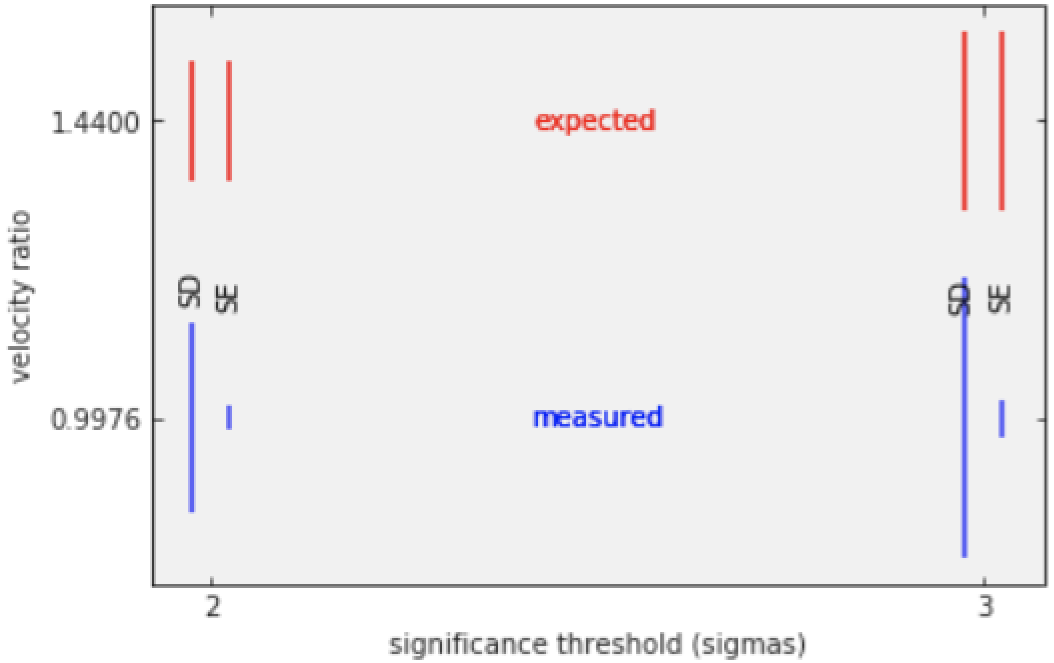
Precision vs Accuracy

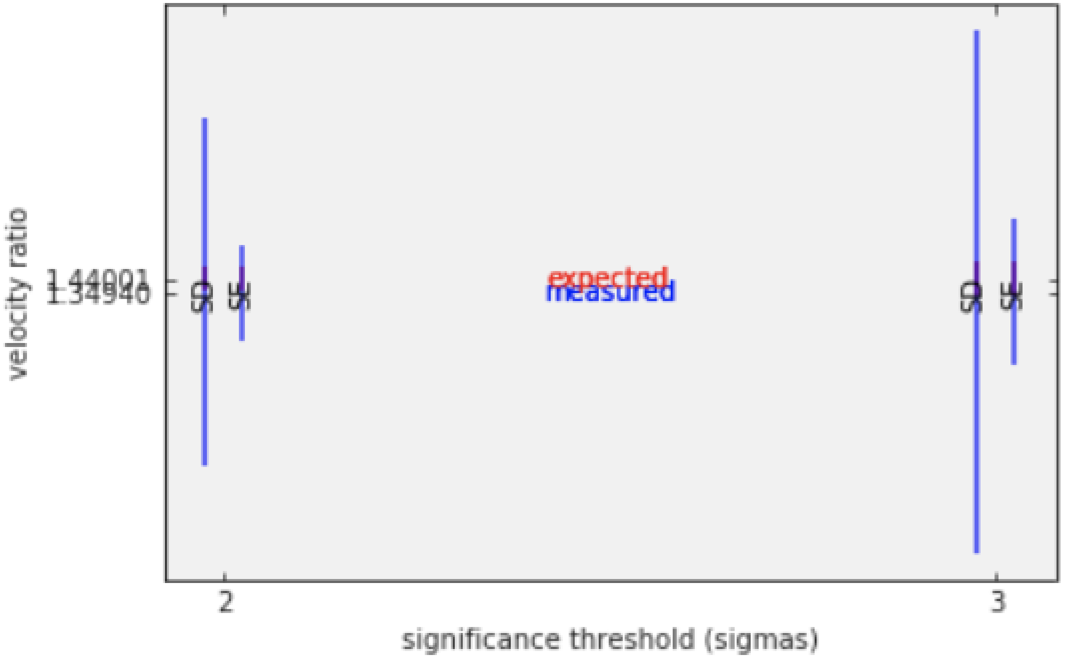
recap




v
radius
2xradius
3xradius
Velocity (km/s)
radius (kpc)
240.2
(74.9, 20.5)
178.0
(47.0, 16.3)
179.7
(62.6, 20.9)


v
radius
2xradius
3xradius
Velocity (km/s)
radius (kpc)
236.5
(11.7, 1.6)
237.1
(109.4, 23.9)
218.5
(132.4, 24.7)


velocity ratio
1.4400
0.9976
expected
measured
SD
SE
SD
SE
significance threshold (sigmas)


velocity ratio
1.44001
1.34940
expected
measured
SD
SE
SD
SE
significance threshold (sigmas)

$$z = x \pm y \sim| \quad \sim dz = \sqrt{dx^2 + dy^2} \tag{1}$$

$$z = x * y \sim| \quad \sim dz = |xy| \sqrt{\left(\frac{dx}{x}\right)^2 + \left(\frac{dy}{y}\right)^2} \tag{2}$$

$$z = x / y \sim| \quad \sim dz = \left|\frac{x}{y}\right| \sqrt{\left(\frac{dx}{x}\right)^2 + \left(\frac{dy}{y}\right)^2} \tag{3}$$

$$z = x^n \sim| \quad \sim dz = |n| \sim x^{n-1} dx \tag{4}$$

$$z = cx \sim| \quad \sim dz = |c| \sim dx \tag{5}$$

$$z = f(x, y) \sim| \sim dz = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 dx^2 + \left(\frac{\partial f}{\partial y}\right)^2 dy^2} \tag{6}$$

1 what is a model

2 the principle of parsimony

epistemology

3 fitting a model to data

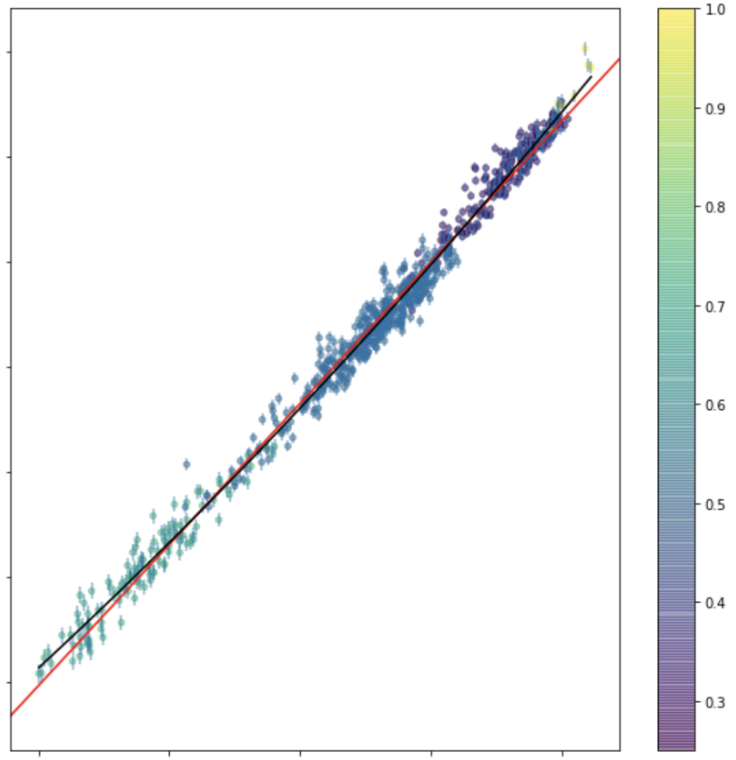
line fit

standard linear fit

higher order equation

uncertainties in the fit parameters

generative models



4 cross-validation

0

what is machine learning?

what is machine learning?

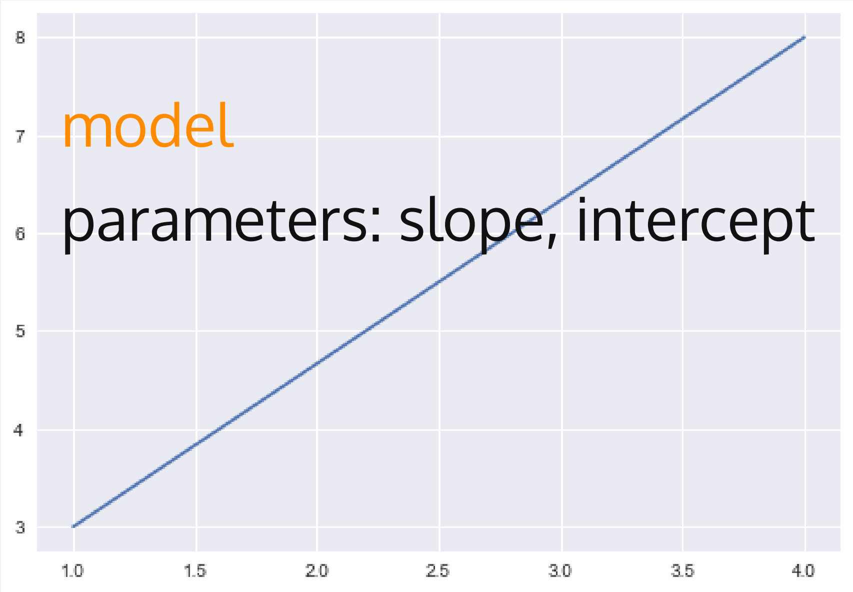
[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

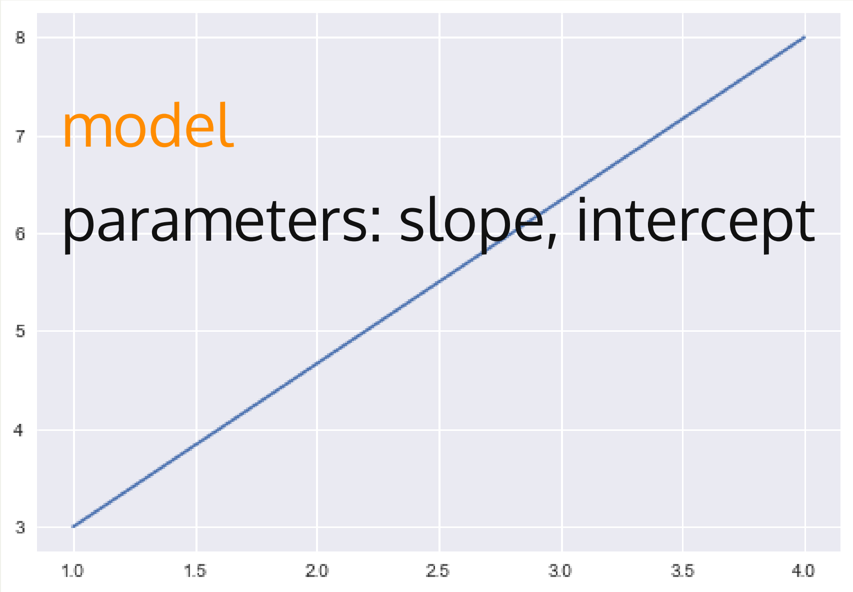
Arthur Samuel, 1959

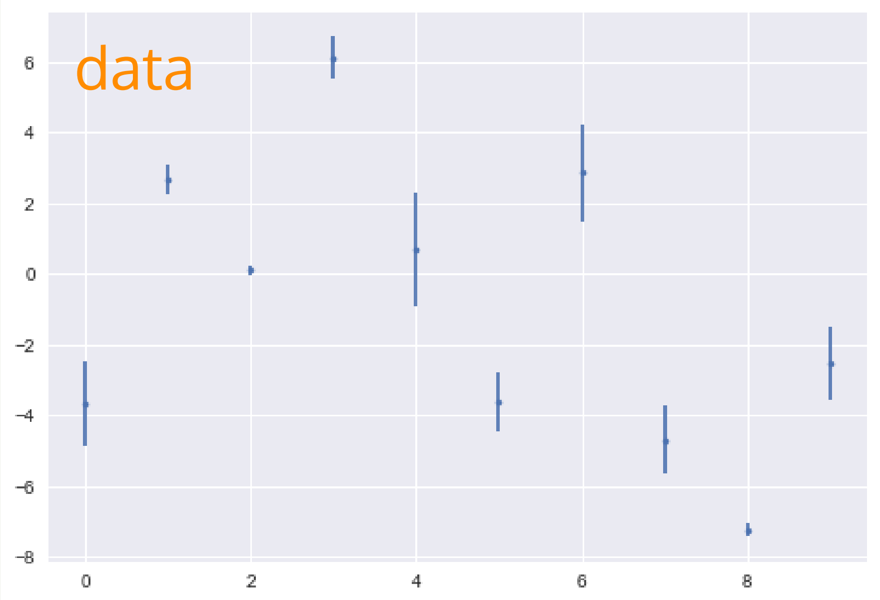


what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

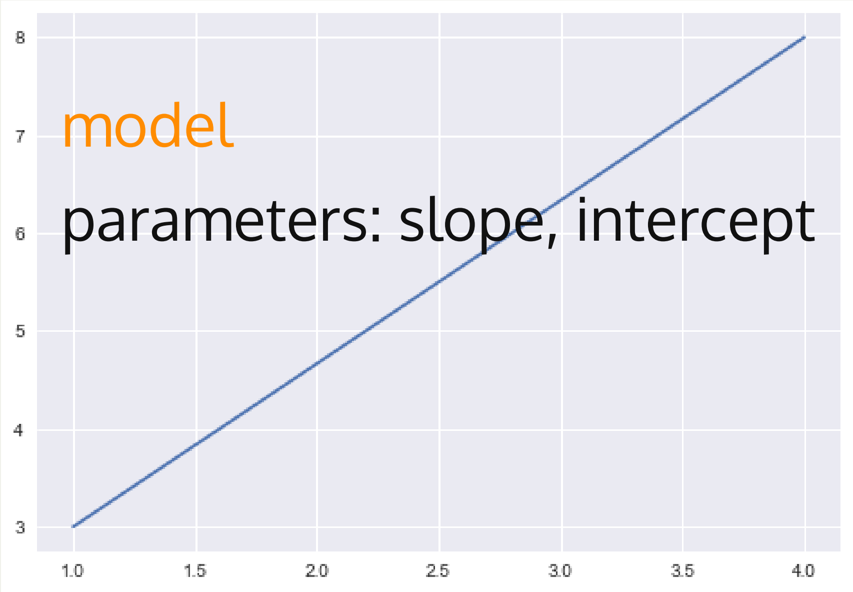




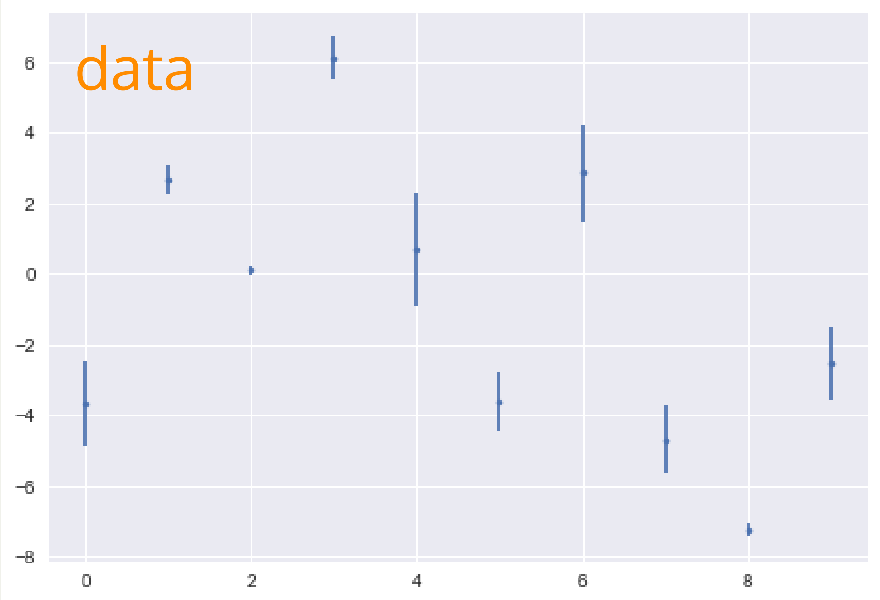
what is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959



ML: any model with parameters learnt from the data



what is machine learning?

supervised learning

extract features and create models that allow prediction where the correct answer is known for a subset of the data

unsupervised learning

identify features and create models that allow to understand structure in the data

what is machine learning?

- k-Nearest Neighbors
- Regression
- Support Vector Machines
- Neural networks
- Classification/Regression Trees

supervised learning

extract features and create models that allow prediction where the correct answer is known for a subset of the data

- clustering
- Principle Component Analsysis
- Apriori (association rule)

unsupervised learning

identify features and create models that allow to understand structure in the data

what is machine learning?

- k-Nearest Neighbors
- Regression
- Support Vector Machines
- Neural networks
- Classification/Regression Trees

supervised learning

classification

prediction

feature selection

- clustering
- Principle Component Analsysis
- Apriori (association rule)

unsupervised learning

understanding structure

organizing + compressing data

anomaly detection

dimensionality reduction

1 what is a model?

what is a model?

it's always a simplification

what is a model?

a *mathematical* representastion of reality

In applying mathematics to subjects such as physics or statistics we make tentative assumptions about the real world which we know are false but which we believe may be useful nonetheless.

George Box, 1976

- no model is right
- some models are useful

what is a model?

why do we model?

1

https://www.youtube.com/embed/Tk2v1UaTgmk?enablejsapi=1

1. to undestand

why do we model?

1

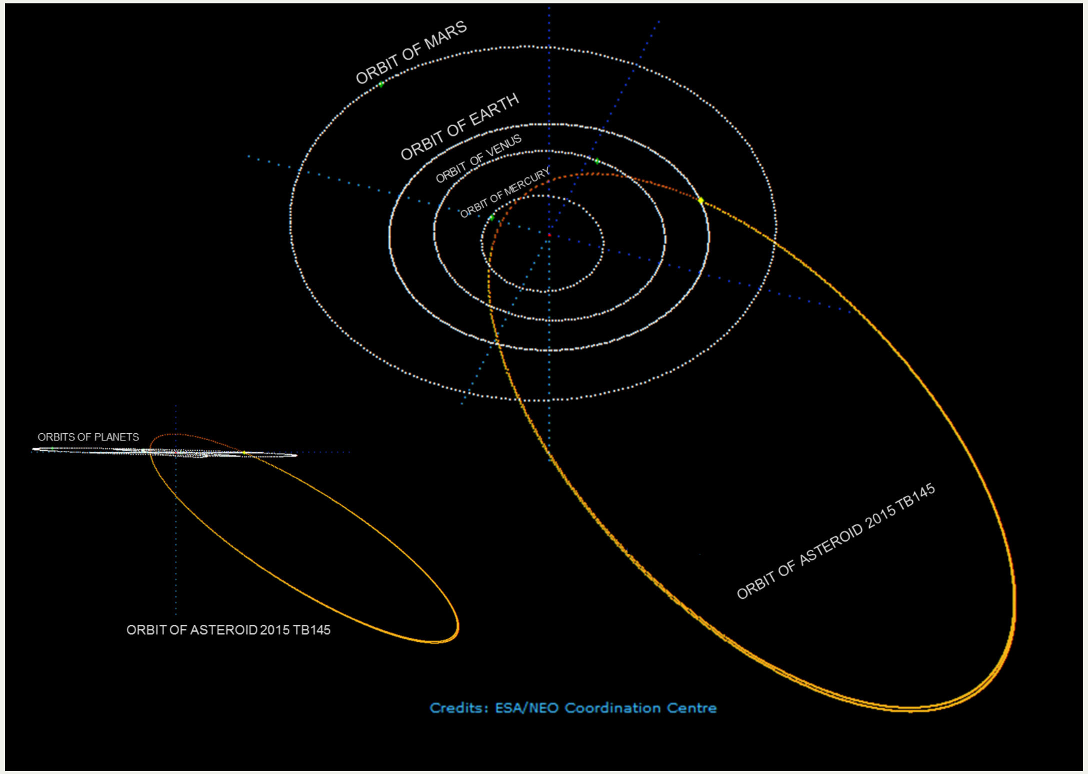
https://www.youtube.com/embed/Tk2v1UaTgmk?enablejsapi=1

1. to undestand

$$h(t+t_0) = v_0 t - \frac{1}{2} g t^2$$

what is a model?

why do we model?

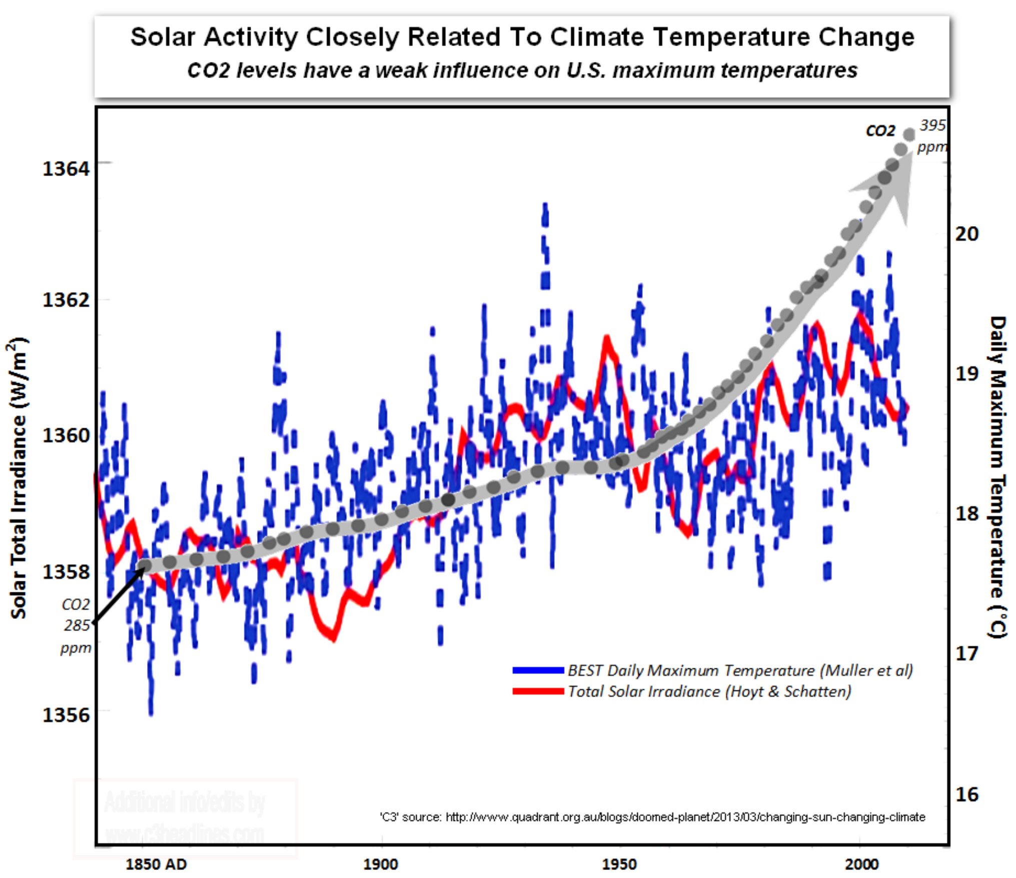


2. to predict

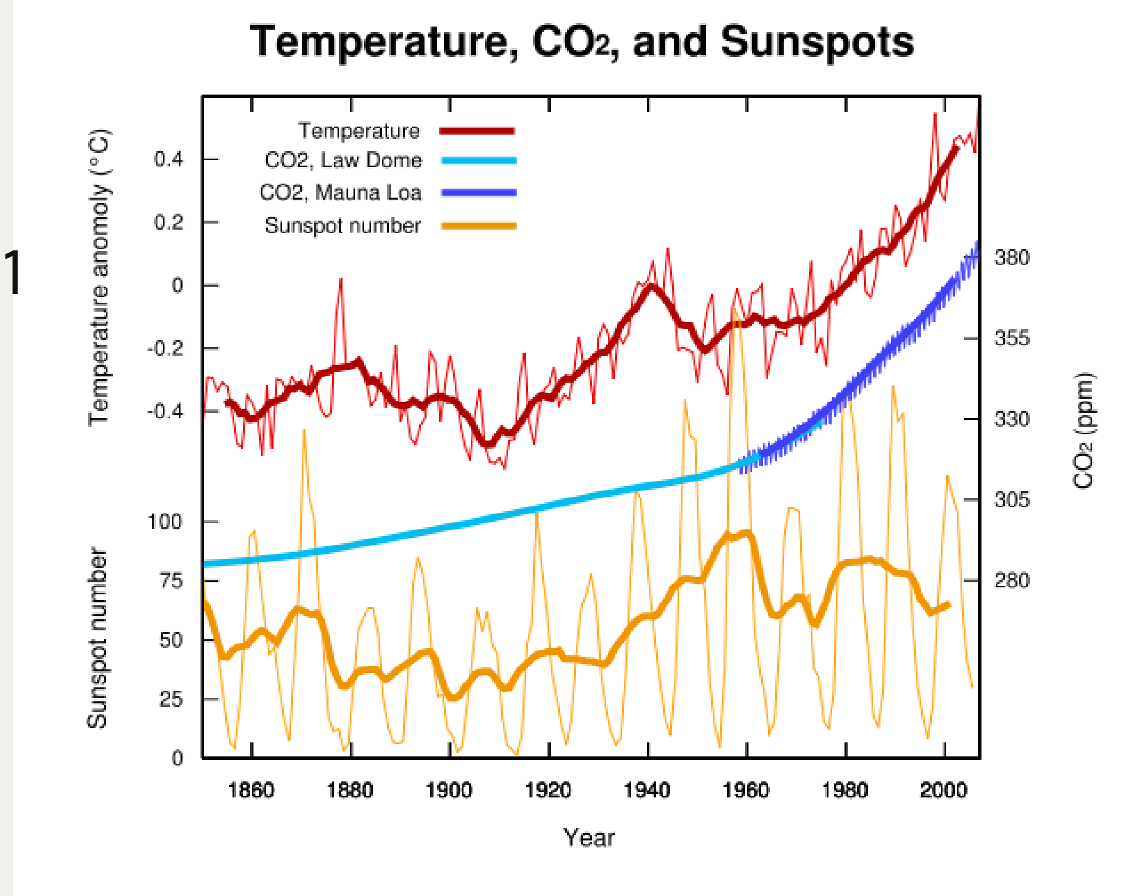
$$r(\nu) = \frac{a(1-e^2)}{1+e\cos(\nu)}$$

what is a model?

why do we model?



1



to understand + predict

what is a model?

the best way to think about it *in the ML context:*

a model is a low dimensional representation of a higher dimensionality datase

2 the principle of parsimony

the princile of parsimony

or Ockham's razor

Pluralitas non est ponenda sine neccesitate

William of Ockham (logician and Franciscan friar) 1300ca

but probably to be attributed to John Duns Scotus (1265–1308)

“Complexity needs not to be postulated without a need for it”

the princile of parsimony

the earth is round,

and it orbits around the sun

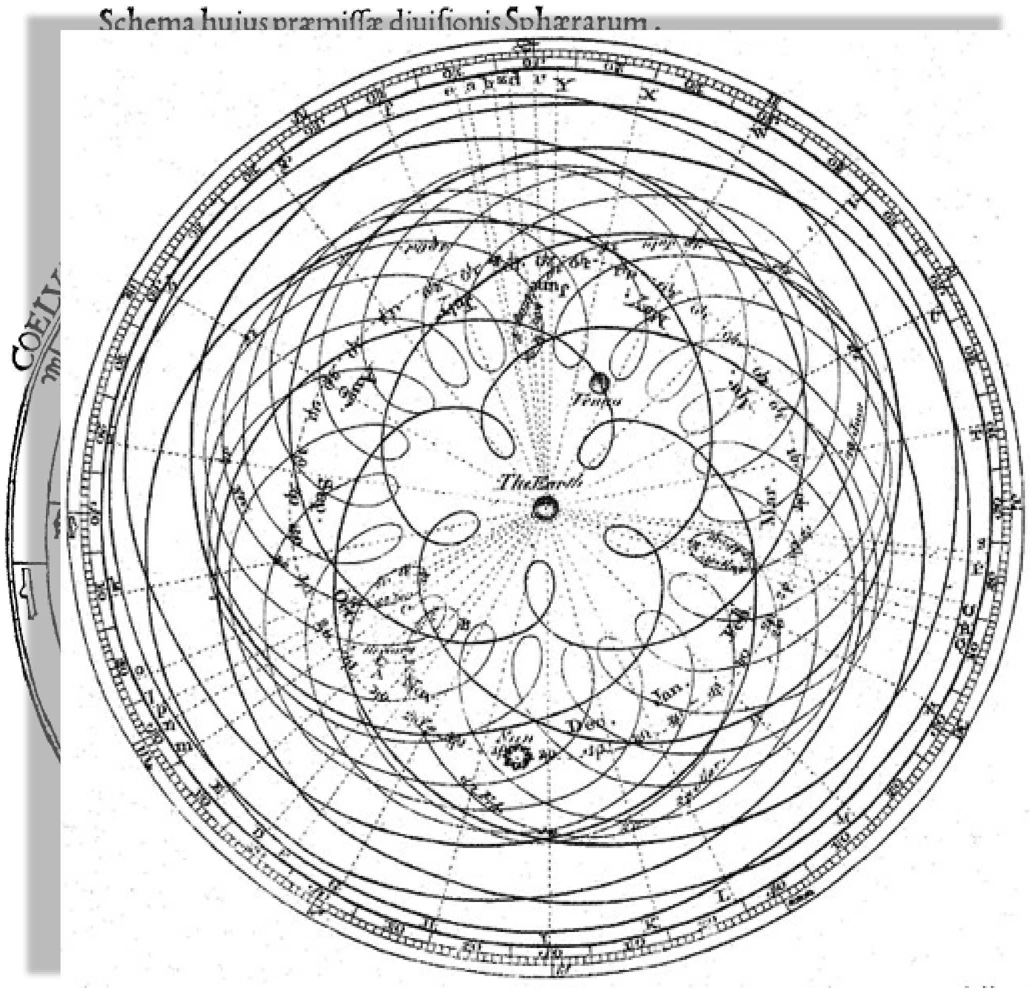
Geocentric models are intuitive:

from our perspective we see the Sun moving, while we stay still

Peter Apian, *Cosmographia*, Antwerp, 1524 from Edward Grant,

"Celestial Orbs in the Latin Middle Ages", *Isis*, Vol. 78, No. 2. (Jun., 1987).

the princile of parsimony

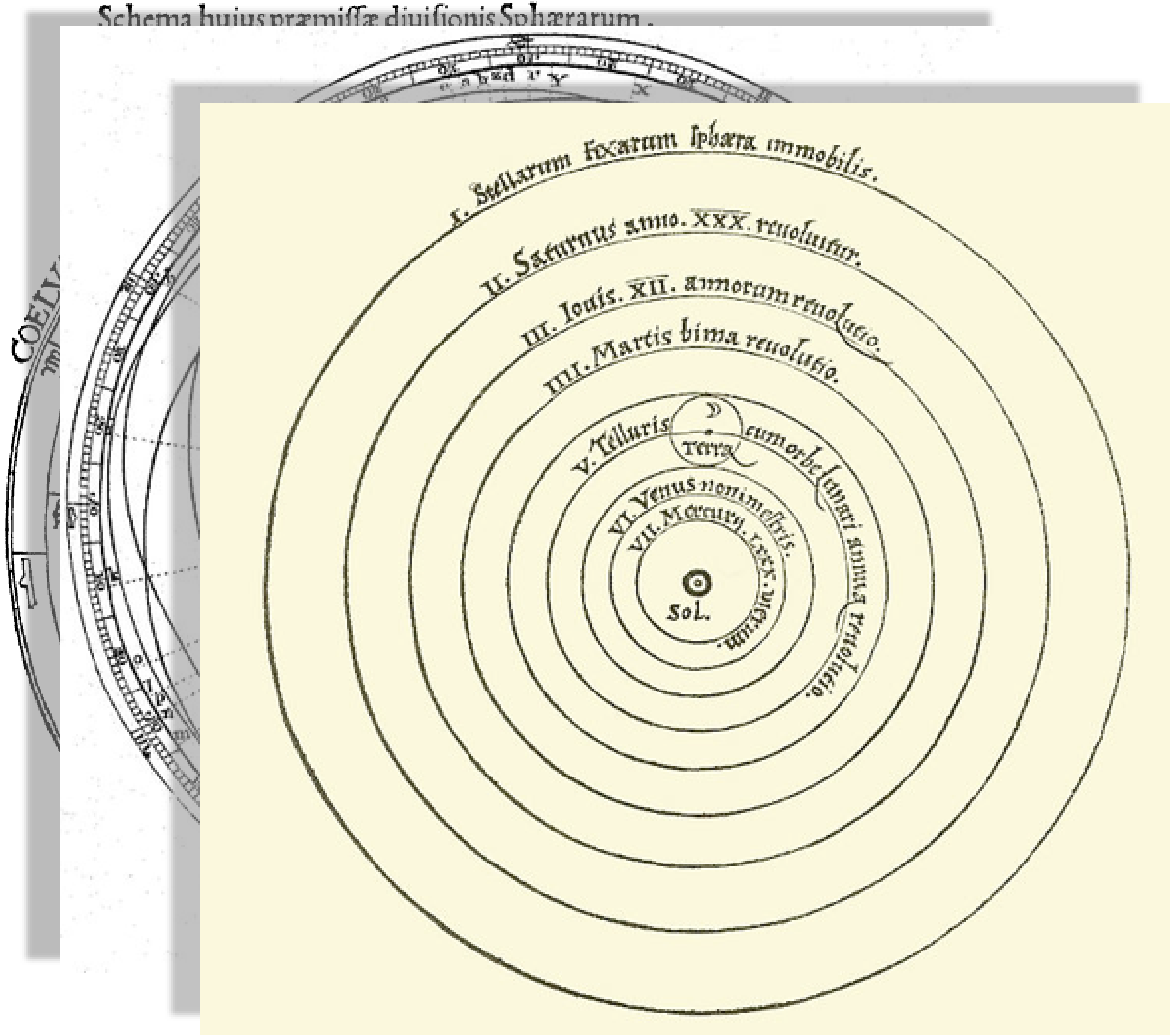


Encyclopaedia Brittanica 1st Edition
Dr Long's copy of Cassini, 1777

the earth is round,
and it orbits around the sun

As observations improve
this model can no longer fit the data!
not easily anyways...

the princile of parsimony



the earth is round,

~~and it orbits around the sun~~

A new model that is much simpler fit the data just as well

(perhaps though only until better data comes...)

Heliocentric model from Nicolaus Copernicus' *De revolutionibus orbium coelestium.*

the princile of parsimony

or Ockham's razor

Pluralitas non est ponenda sine neccesitate

William of Ockham (logician and Franciscan friar) 1300ca
but probably to be attributed to John Duns Scotus (1265–1308)

“Complexity needs not to be postulated without a need for it”

the princile of parsimony

or Ockham's razor

Pluralitas non est ponenda sine neccesitate

William of Ockham (logician and Franciscan friar) 1300ca
but probably to be attributed to John Duns Scotus (1265–1308)

“Complexity needs not to be postulated without a need for it”
“Between 2 theories that perform similarly choose the ***simpler one***”

the princile of parsimony

or Ockham's razor

Pluralitas non est ponenda sine neccesitate

William of Ockham (logician and Franciscan friar) 1300ca
but probably to be attributed to John Duns Scotus (1265–1308)

“Complexity needs not to be postulated without a need for it”
“Between 2 theories that perform similarly choose the *one with fewer parameters*”

the princile of parsimony

Science and Statistics George E. P. Box (1976)

Journal of the American Statistical Association, Vol. 71, No. 356, pp. 791-799.

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena

Since all models are wrong the scientist must be alert to what is importantly wrong.

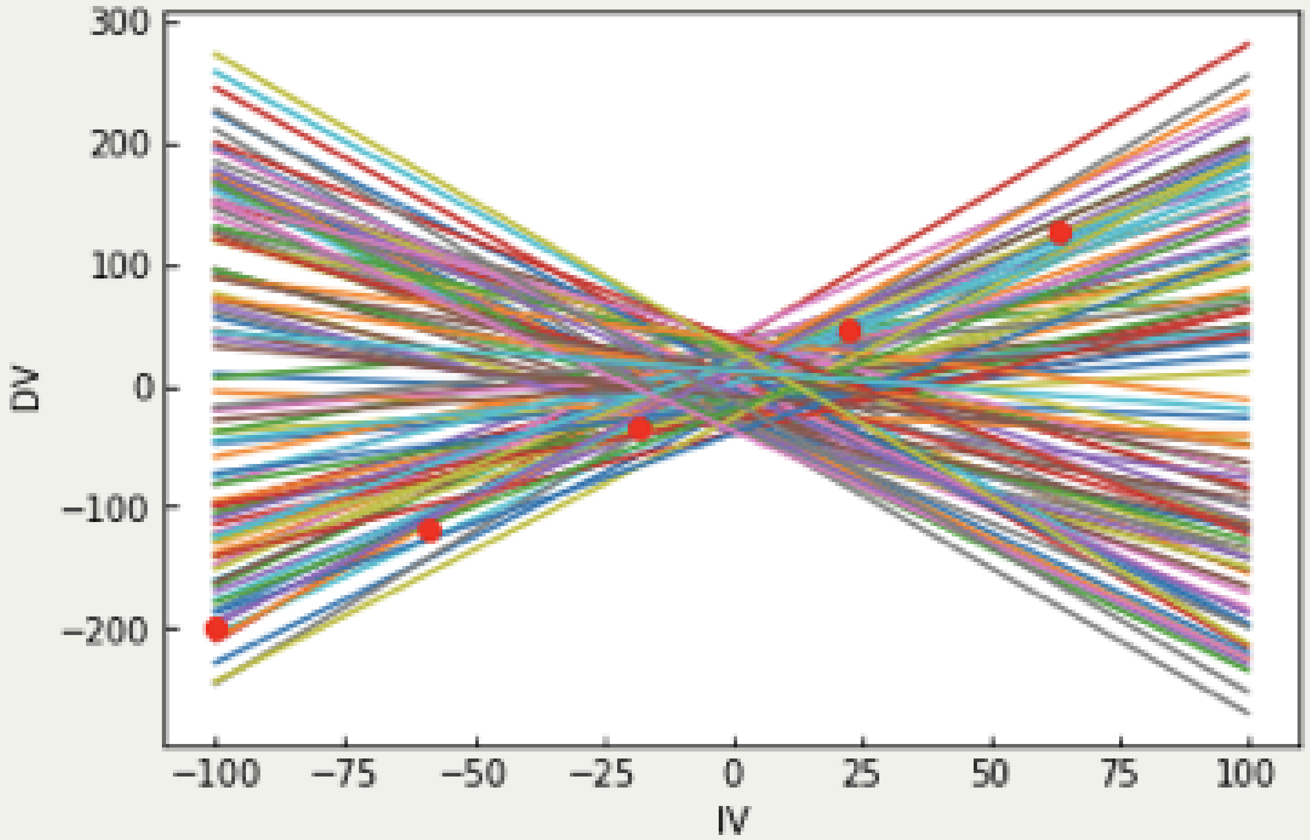
3 fitting a simple model to data

Many packages do it in python!

- numpy (e.g. np.polyfit)
- scipy (e.g. sp.optimize.curve_fit)
- skitlearn.linear_model.LinearRegression()

but what are they doing??

line model: *ax+b*

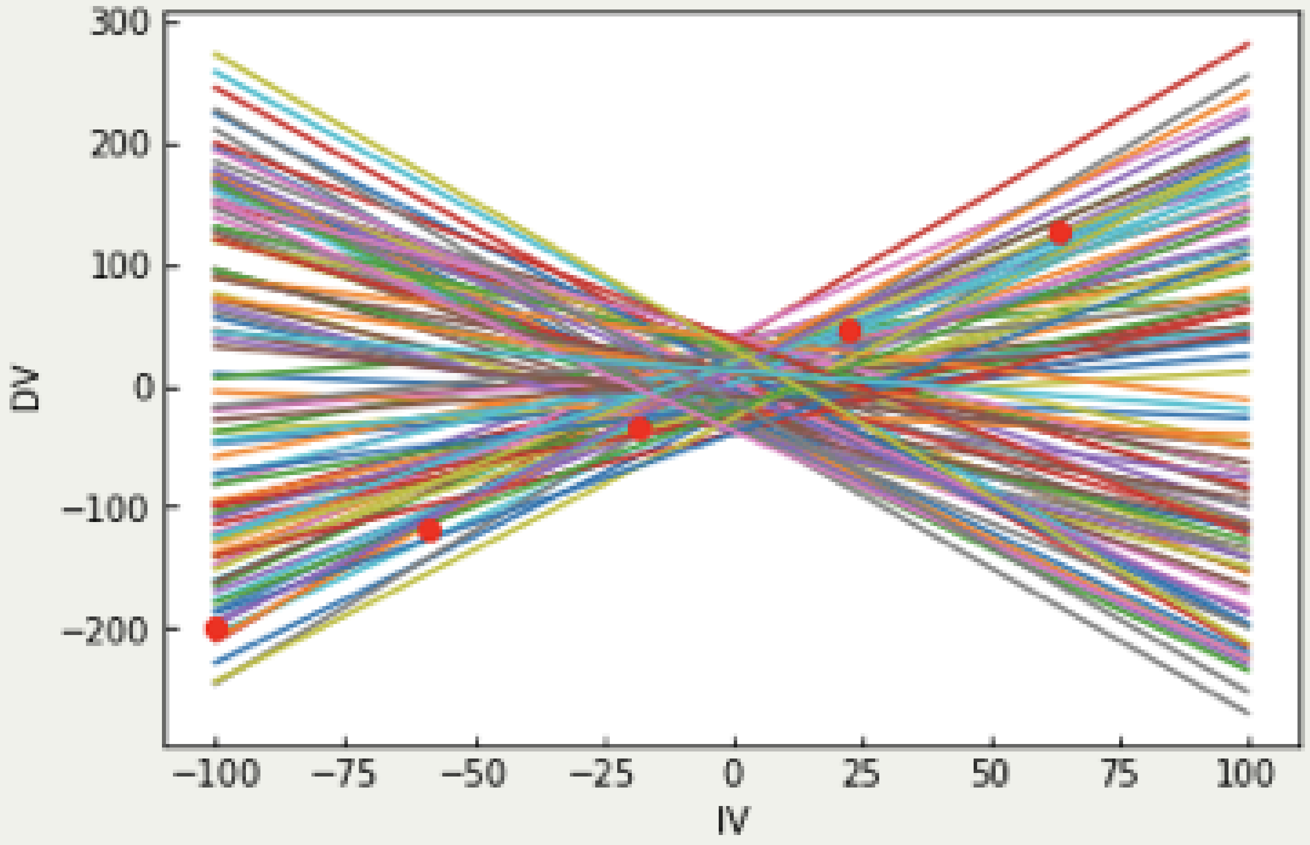


1

choose your model :

choose a mathematical formula to represent the behavior you see/expect in the data

line model: *ax+b*



2

choose an objective function :

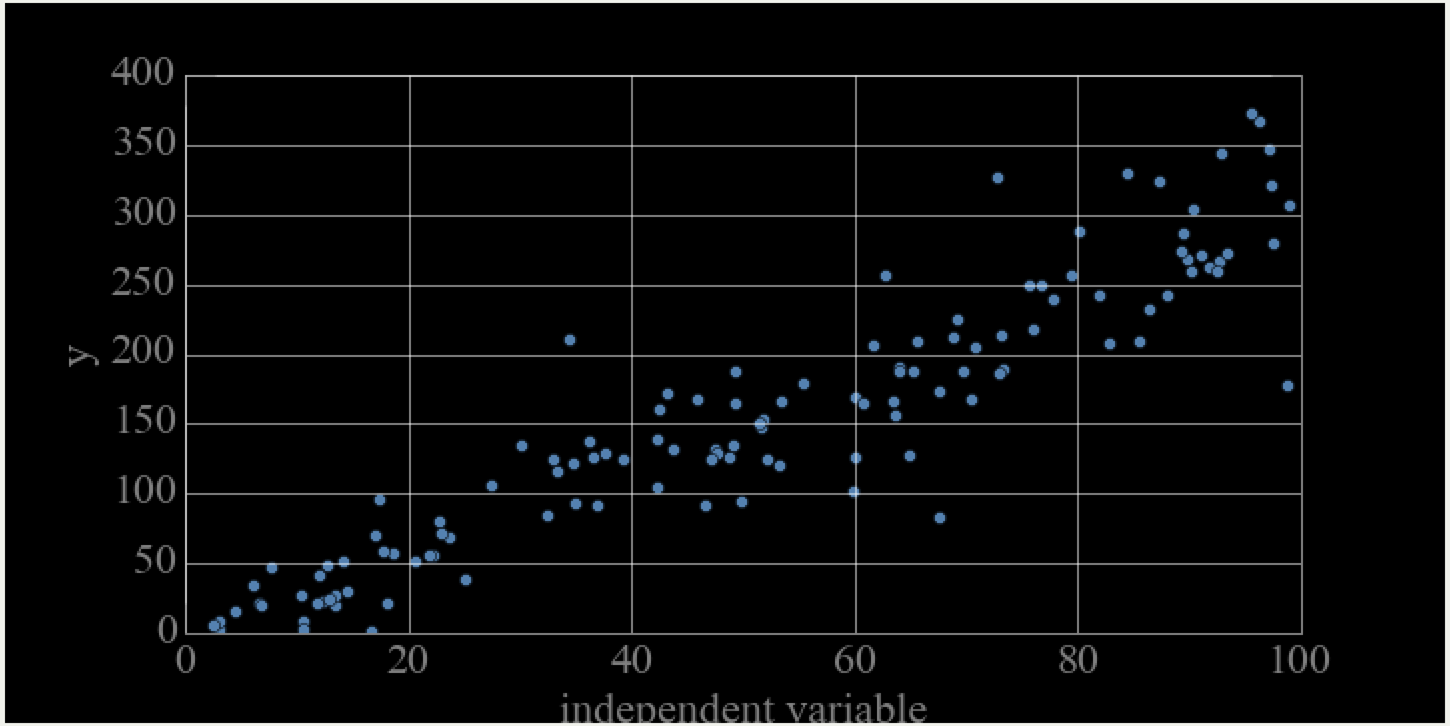
you need a plan to choose the parameters of the model: to "optimize" the model.
You need to choose something to be MINIMIZED or MAXIMIZED

objective function:

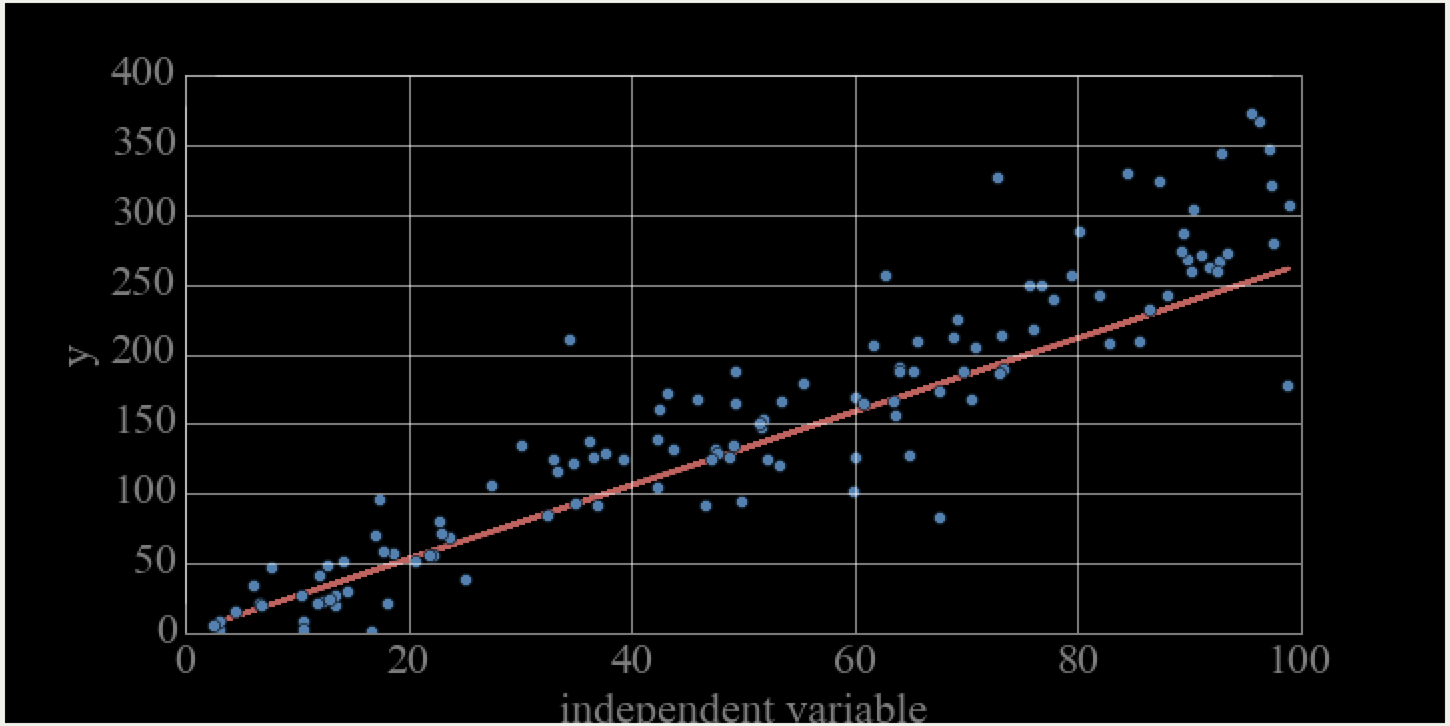
what you want to optimize for

In principle, there are many choices for objective function. But the only procedure that is truly justified—in the sense that it leads to interpretable probabilistic inference, is to make a generative model for the data.

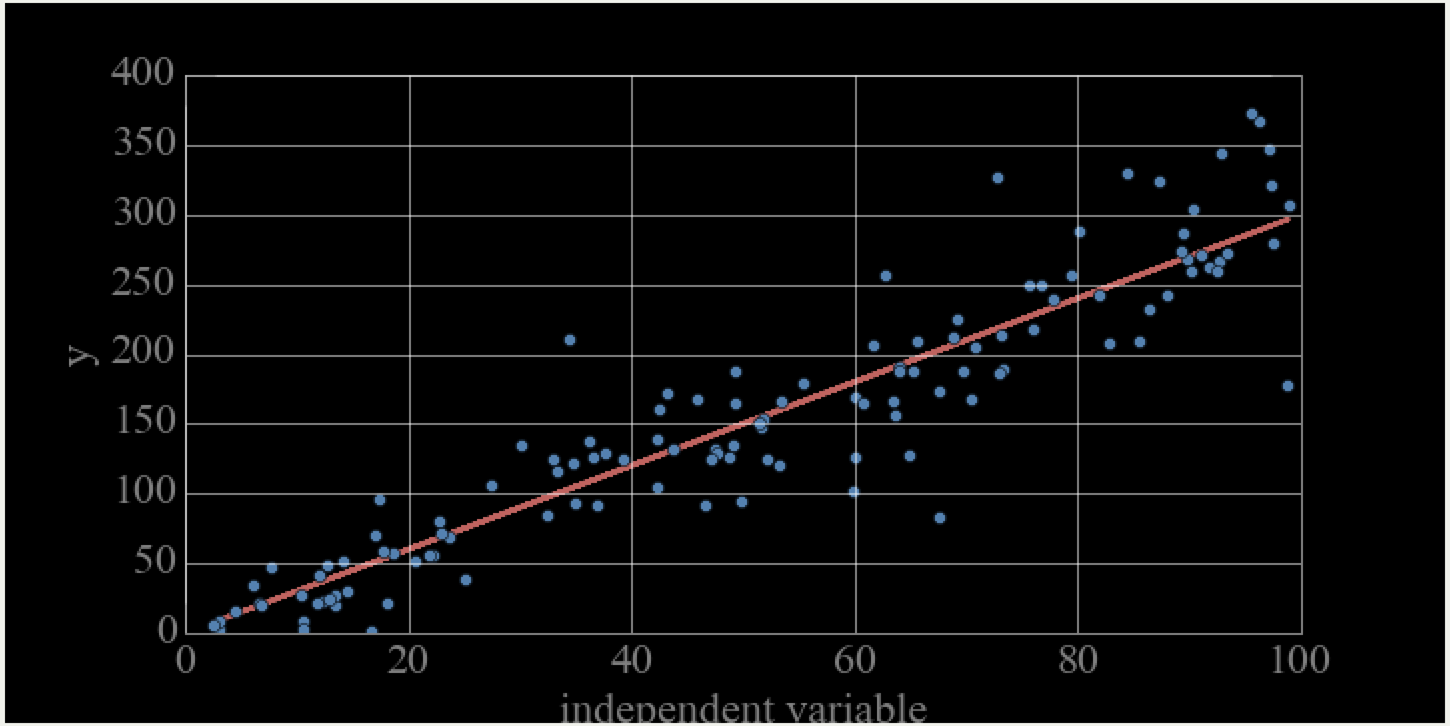
objective function:



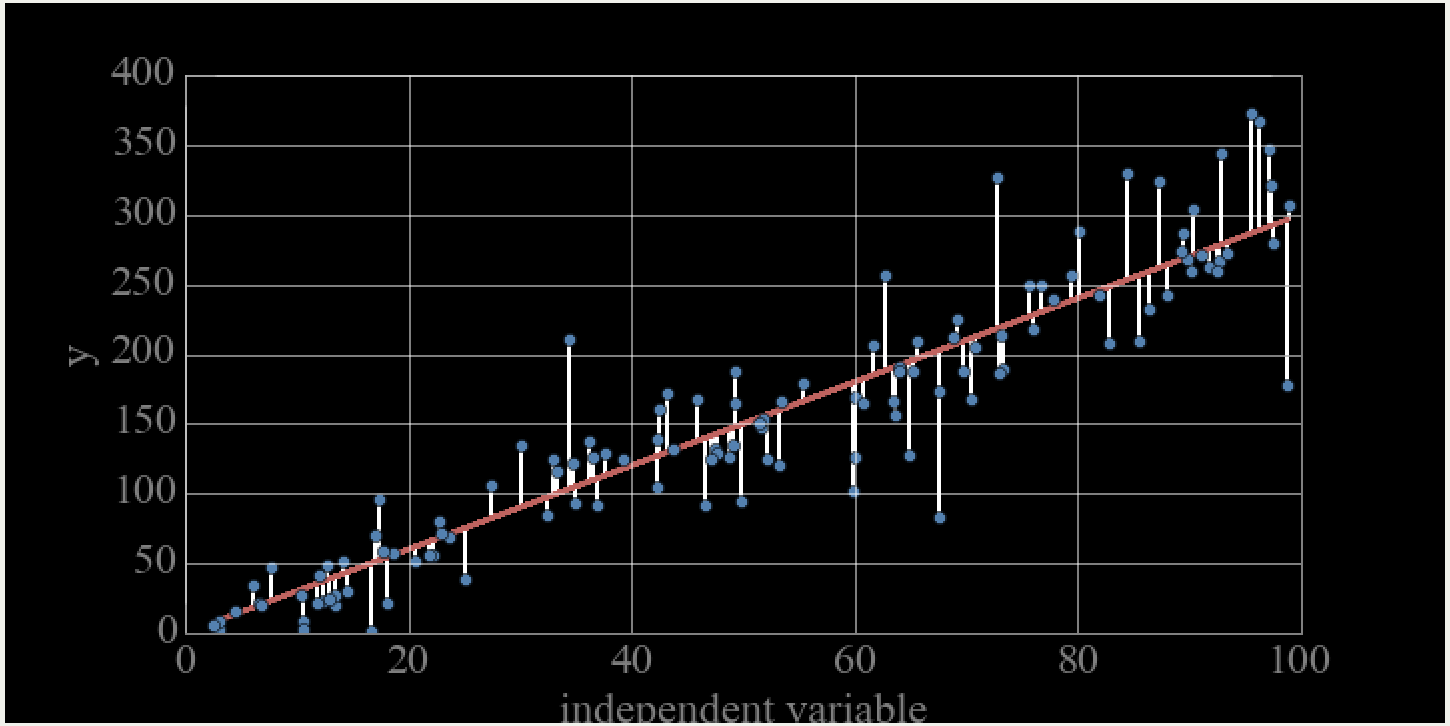
objective function:



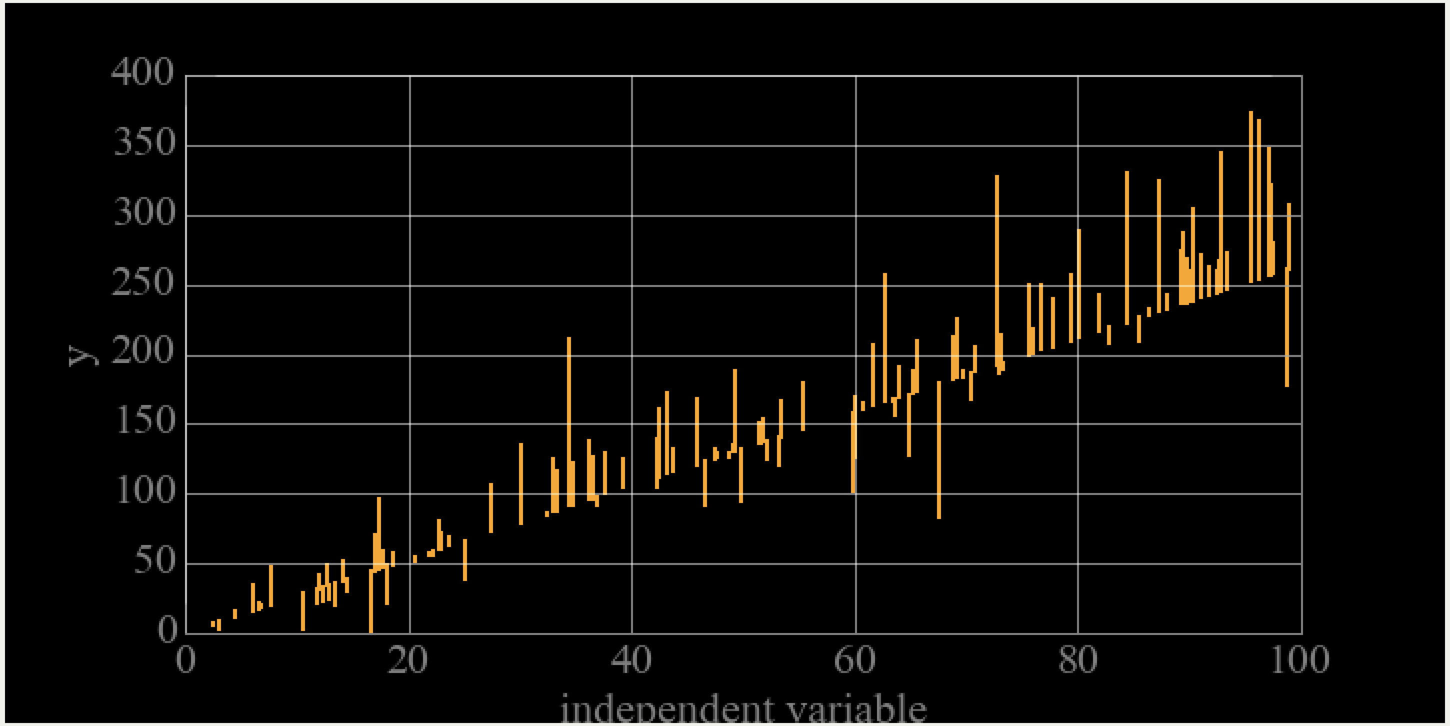
objective function:



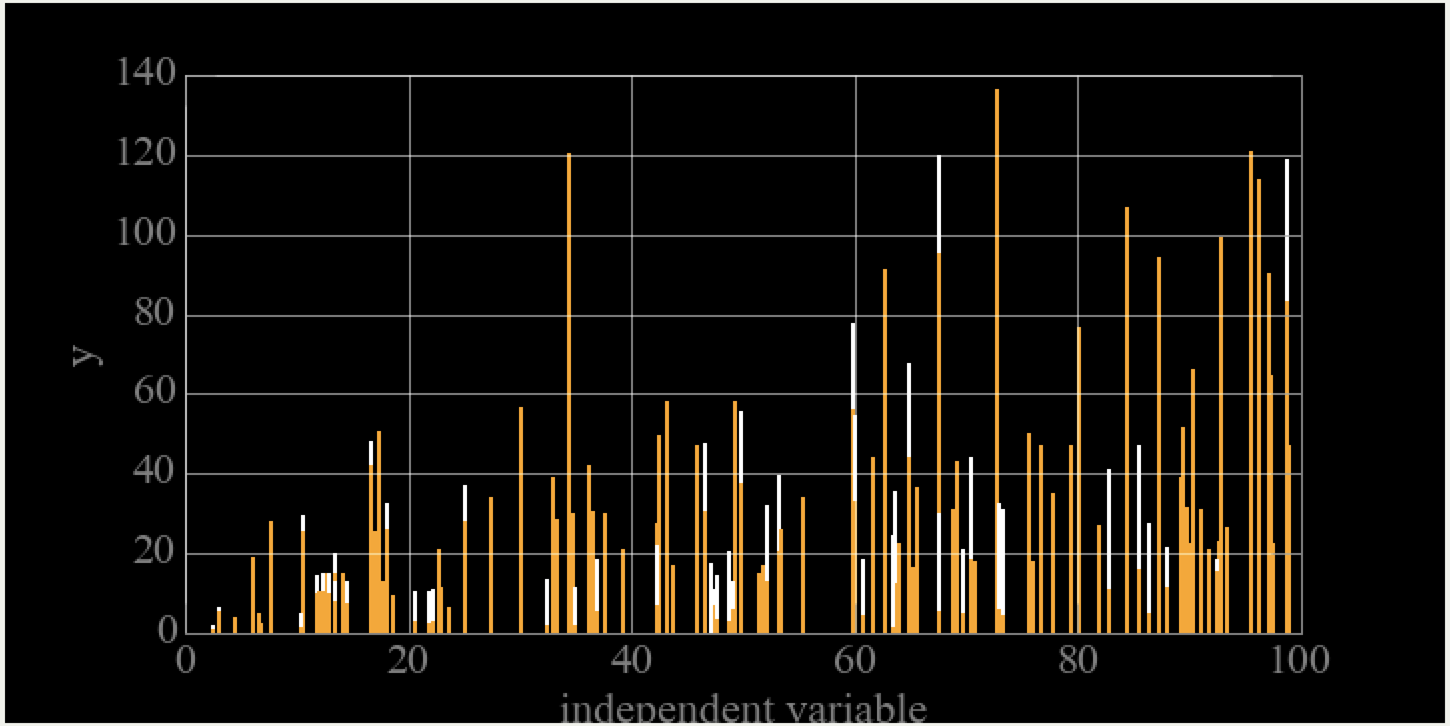
objective function:



objective function:



objective function:



$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

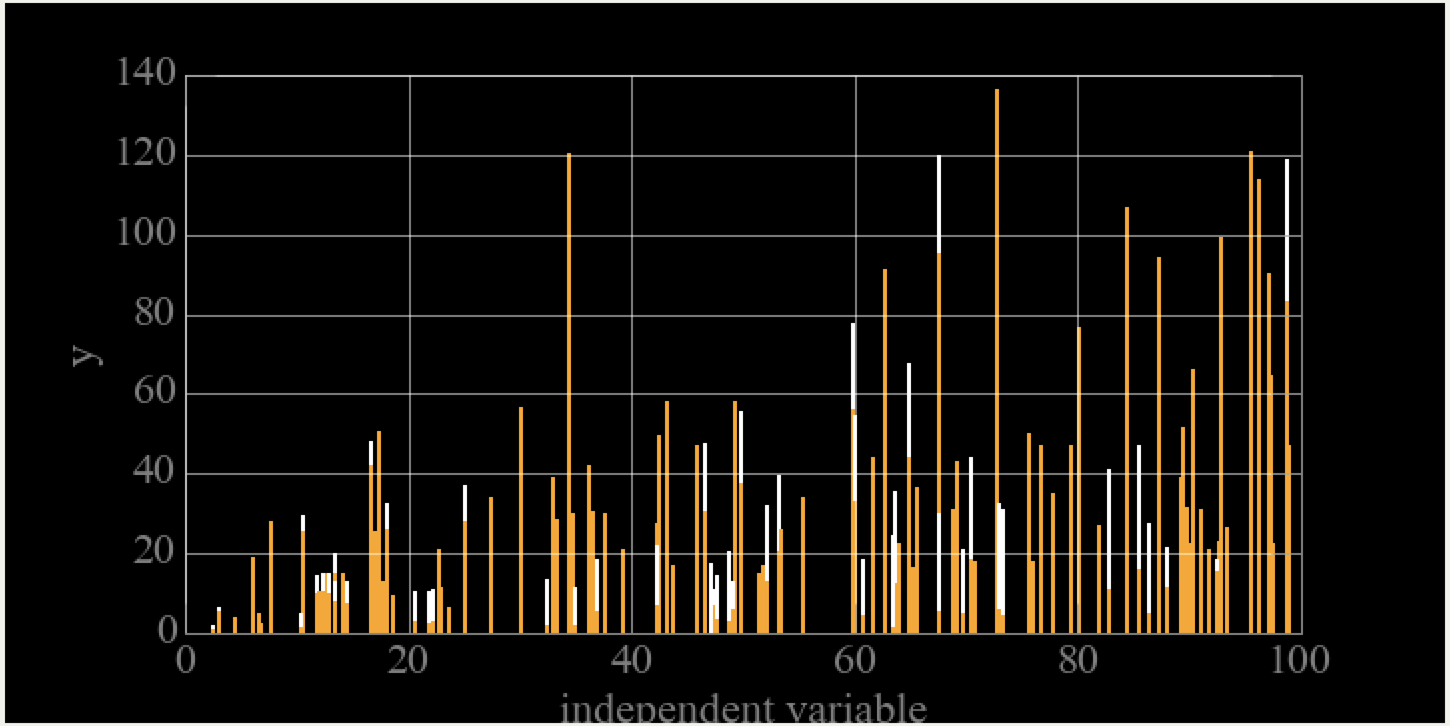
yi: i-th observation

xi: i-th measurement "location"

si: i-th uncertainty

Fit model parameters <==> minimize the Sum of residuals squared

objective function:



$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

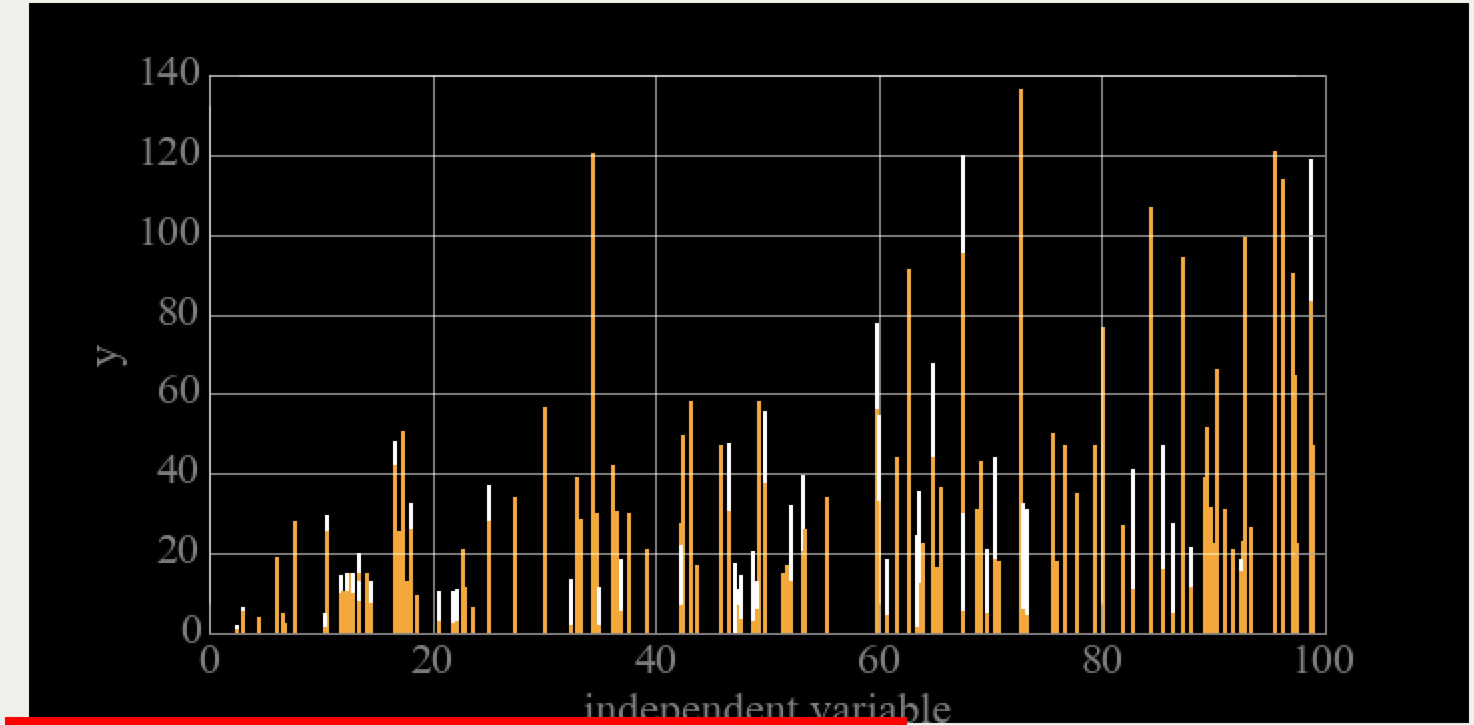
yi: i-th observation

fi: i-th prediction

si: i-th uncertainty

Fit model parameters <==> minimize the Sum of residuals squared

objective function:



if you have uncertainties

(almost always in observational and experimental phsyics)

$$\boxed{\chi^2 = \sum \frac{(y_i - (mx_i + b))^2}{2\sigma_i^2}}$$

yi: i-th observation

xi: i-th measurement "location"

sigmai: i-th uncertainty

Fit model parameters <==> minimize the Sum of residuals squared

objective function:

what you want to optimize for

homeoscedastic :

the uncertainty is the same for all data points

heteroscedastic :

the uncertainty different for each datapoint

(almost always the case in physics!)

$$\chi^2 = \sum \frac{(y_i - (mx_i + b))^2}{2\sigma_i^2}$$

yi: i-th observation

xi: i-th measurement "location"

sigmai: i-th uncertainty

Fit model parameters <==> minimize the Sum of residuals squared

objective function:

what you want to optimize for

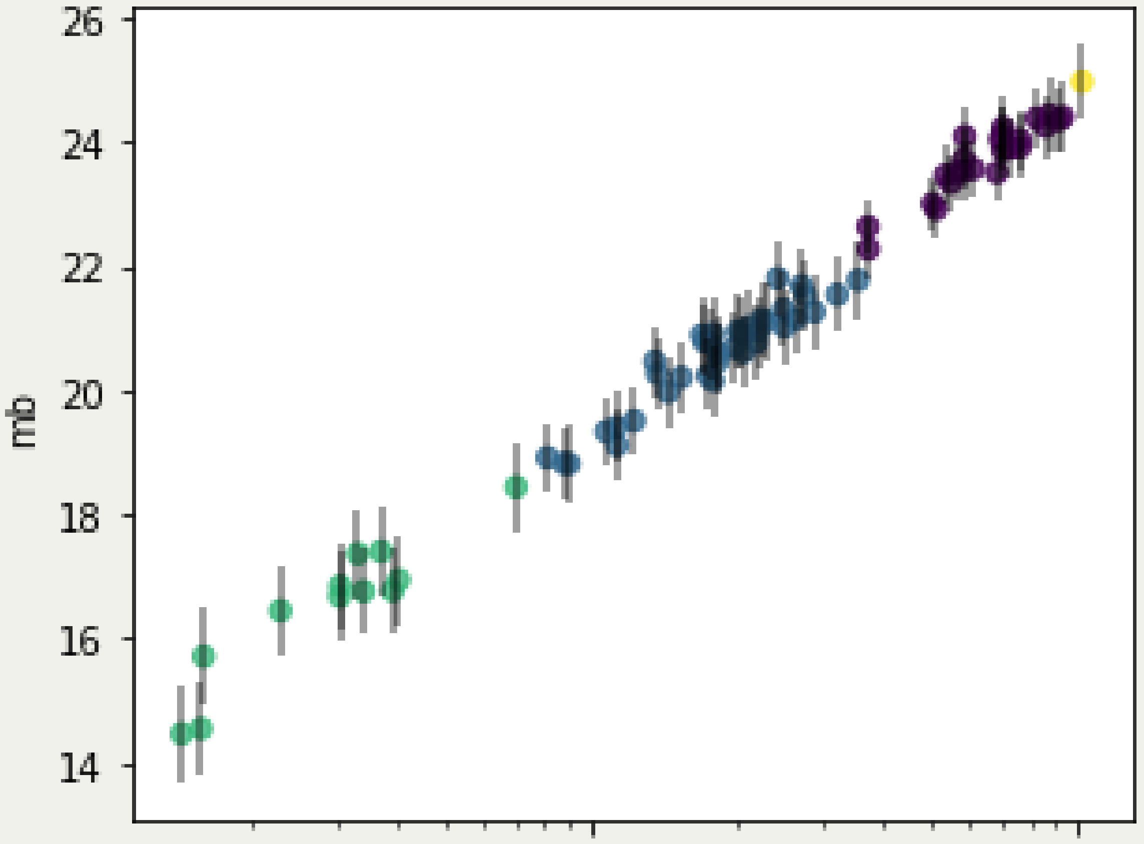
homeoscedastic :

the uncertainty is the same for all data points

heteroscedastic :

the uncertainty different for each datapoint

(almost always the case in physics!)



datapoints have different uncertainty

(almost always in pysyics)

objective function:

what you want to optimize for

homeoscedastic :

the uncertainty is the same for all data points

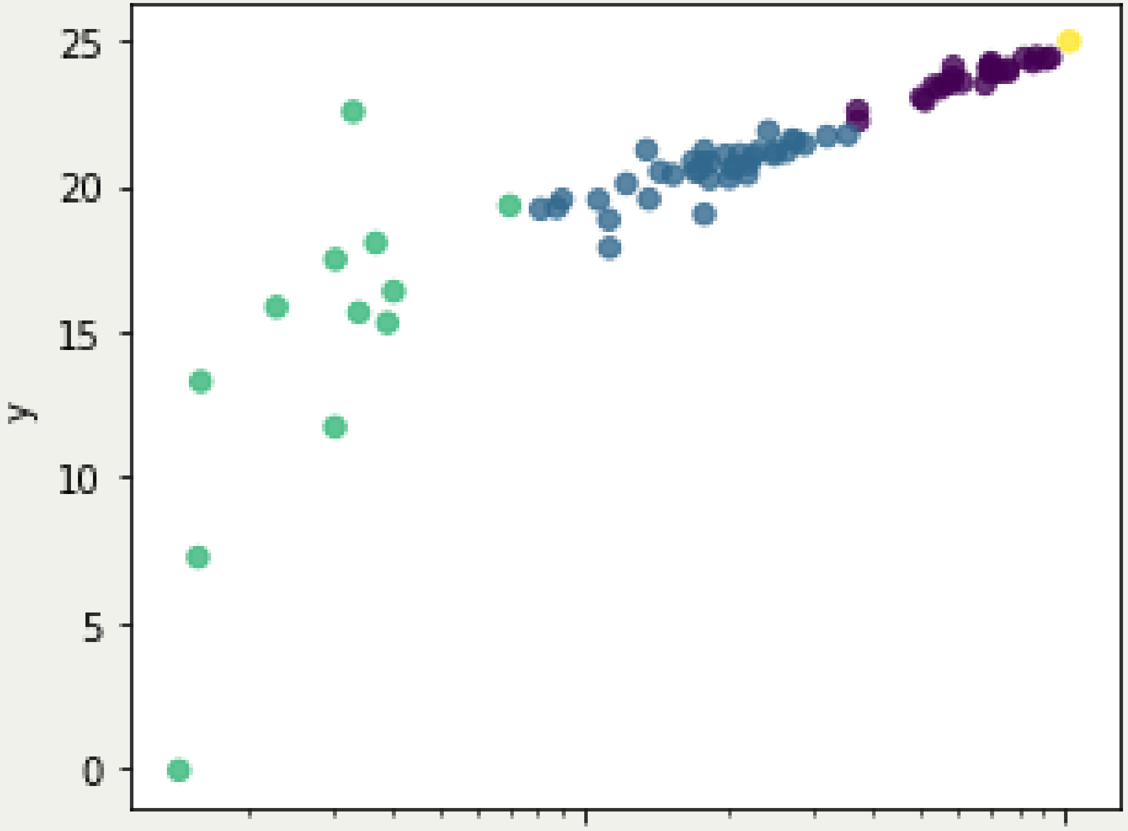
heteroscedastic:

the uncertainty different for each datapoint

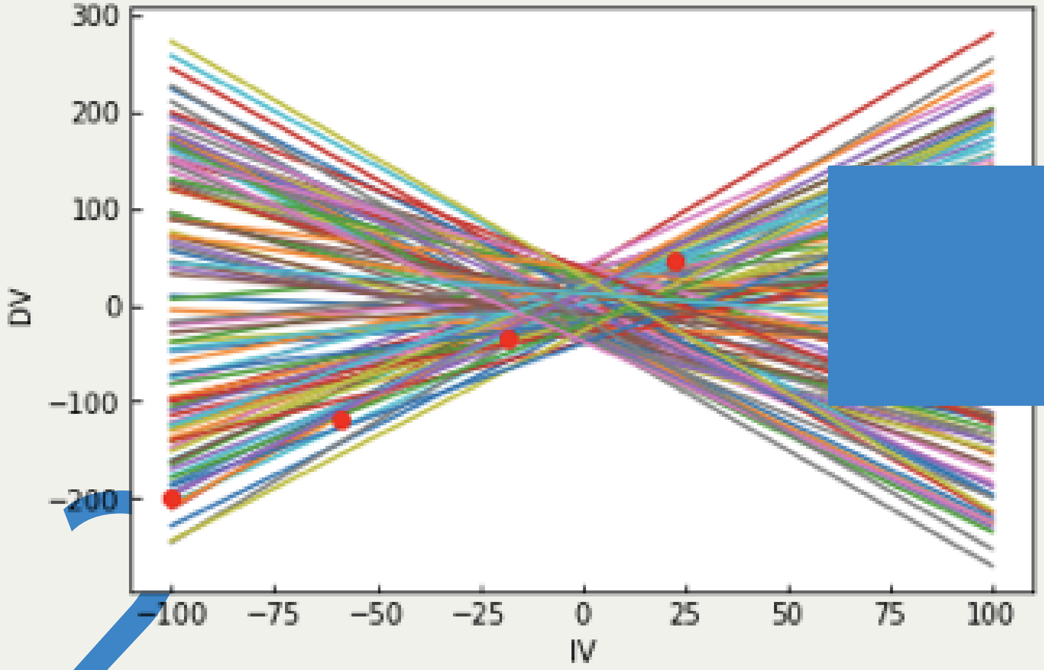
(almost always the case in physics!)

scatter **dependent on exogenous variable**

(very difficult problem not well studied in statistics - very common in physics!)

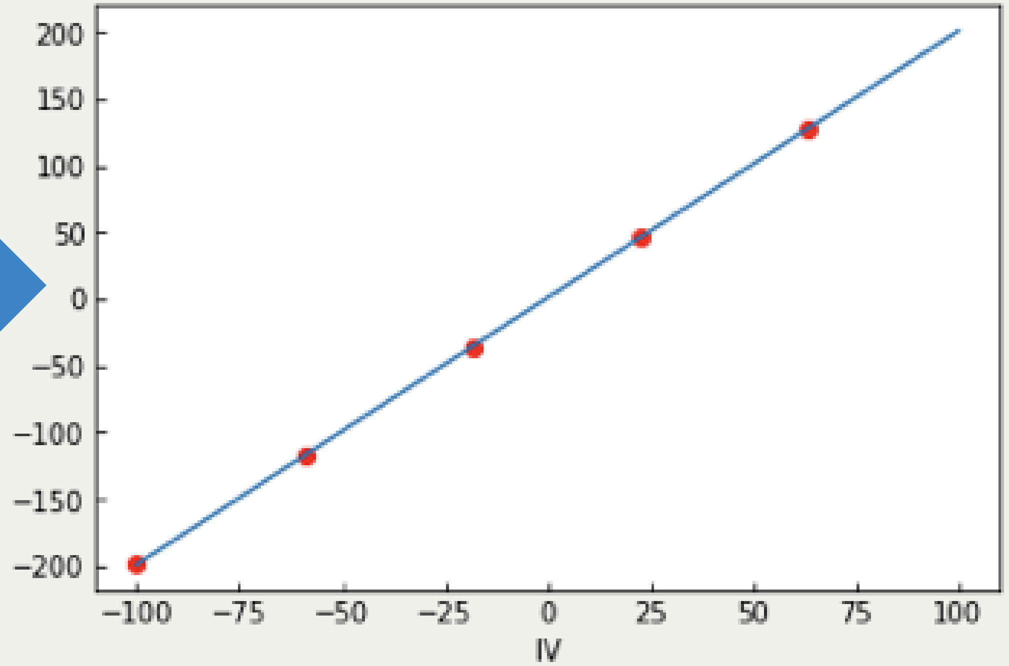


a line is a family of models



line model: *ax+b*

a line with set parameters is a models



2

choose an objective function :

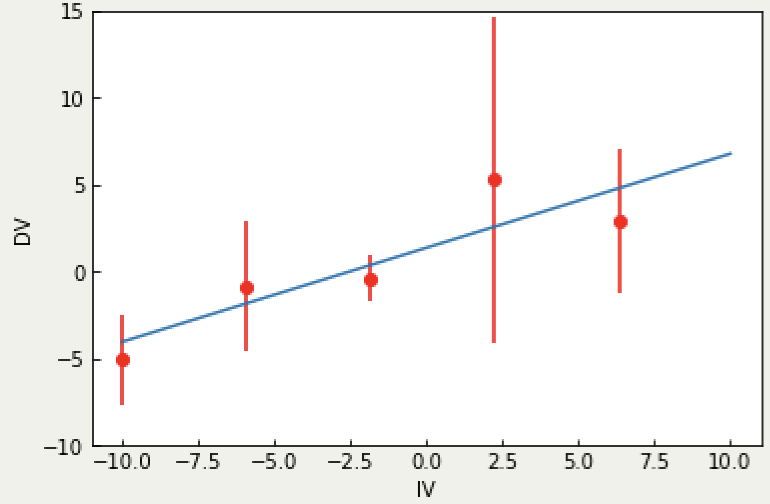
you need a plan to choose the parameters of the model: to "optimize" the model.
You need to choose something to be MINIMIZED or MAXIMIZED

line model: *ax+b*

$$\sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2} \sim \chi^2(dof = DOF)$$

i.e. the X^2 (the quantity above)

follows a X^2 distribution with degrees of freedom equal to the number of degrees of freedom in the problem (generally Ndatapoints - Nparameters)



3

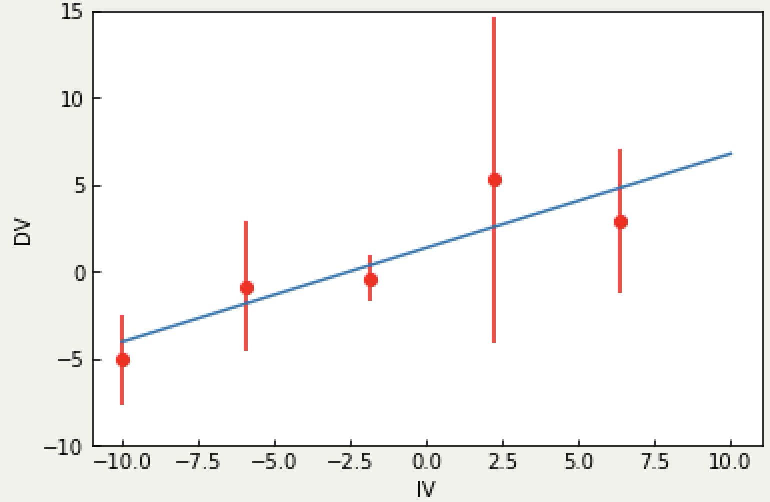
evaluate the quality of your model

again: many options!

line model: *ax+b*

$$\sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2} \sim \chi^2(dof = DOF)$$

$$\frac{\sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}}{DOF} \sim \chi^2(dof = 1)$$



i.e. the "reduced X^2" (X^2/DOF) follows a X^2 distribution with 1 degree of freedom

3

The expectation value (mean) of such a distribution is 1

evaluate the quality of your model

again: many options!

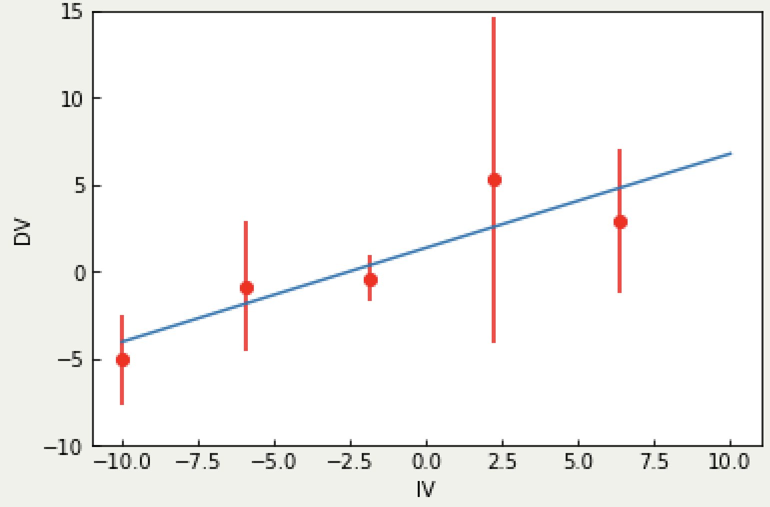
If my model is good I should find

X^2 ~ 1

$$\chi^2_{reduced} = \frac{\sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}}{DOF} \sim 1$$

line model: *ax+b*

$$\sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2} \sim \chi^2(dof = DOF)$$



$$\frac{\sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}}{DOF} \sim \chi^2(dof = 1)$$

i.e. the "reduced X^2" (X^2/DOF) follows a X^2 distribution with 1 degree of freedom

The expectation value (mean) of such a distribution is 1

3

evaluate the quality of your model

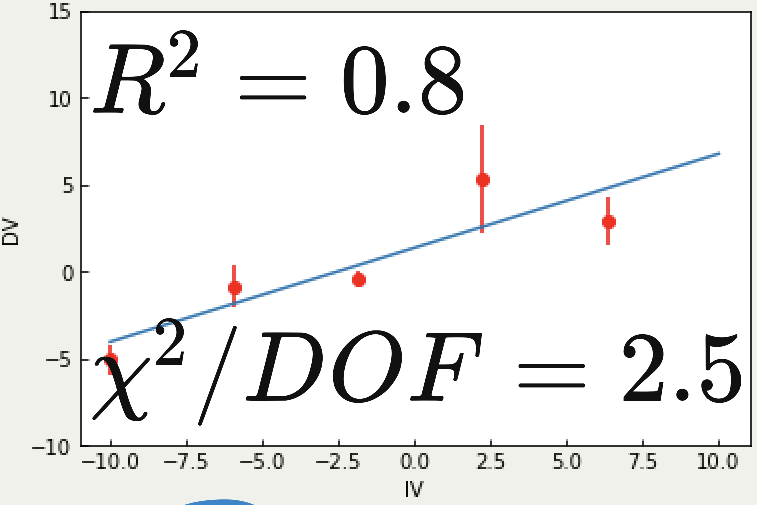
again: many options!

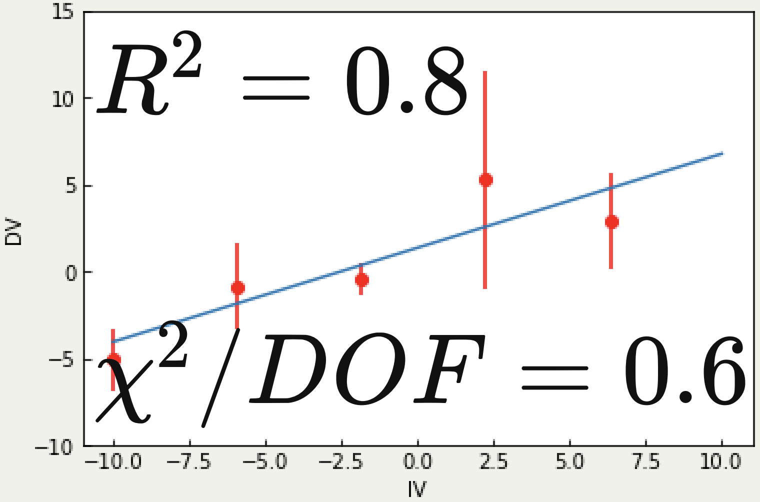
If my model is good I should find

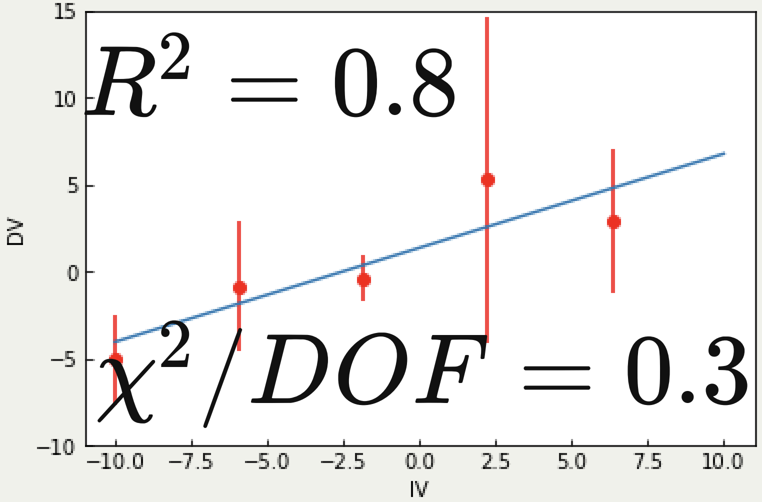
X^2 ~ 1

$$\chi^2_{reduced} = \frac{\sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}}{DOF} \sim 1$$

line model: *ax+b*







3

evaluate the quality of your model

again: many options!

the chi^2 tells you how well your model fits the data given how confident you are in the data (how much you believe they are correct)

The R^2 just tells you about how well the model fits the data

Data analysis recipes:
Fitting a model to data*

David W. Hogg
Center for Cosmology and Particle Physics, Department of Physics, New York University
Max-Planck-Institut für Astronomie, Heidelberg
Jo Bovy
Center for Cosmology and Particle Physics, Department of Physics, New York University
Dustin Lang
Department of Computer Science, University of Toronto
Princeton University Observatory

https://arxiv.org/pdf/1008.4686.pdf

In the case of the straight line fit in the presence of known, Gaussian uncertainties in one dimension, one can create this generative model as follows: Imagine that the data *really do* come from a line of the form $y = f(x) = m\,x + b$, and that the only reason that any data point deviates from this perfect, narrow, straight line is that to each of the true y values a small y-direction offset has been added, where that offset was drawn from a Gaussian distribution of zero mean and known variance σ_y^2. In this model, given an independent position x_i, an uncertainty σ_{yi}, a slope m, and an intercept b, the frequency distribution $p(y_i|x_i,\sigma_{yi},m,b)$ for y_i is

$$p(y_i|x_i,\sigma_{yi},m,b) = \frac{1}{\sqrt{2\,\pi\,\sigma_{yi}^2}}\,\exp\left(-\frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2}\right) \quad , \tag{9}$$

where this gives the expected frequency (in a hypothetical set of repeated experiments[13]) of getting a value in the infinitesimal range $[y_i, y_i + \mathrm{d}y]$ per unit $\mathrm{d}y$.

The generative model provides us with a natural, justified, scalar objective: We seek the line (parameters m and b) that maximize the probability of the observed data given the model or (in standard parlance) the *likelihood of the parameters*.[14] In our generative model the data points are independently drawn (implicitly), so the likelihood $\mathscr{L}$ is the product of conditional probabilities

$$\mathscr{L} = \prod_{i=1}^{N} p(y_i|x_i,\sigma_{yi},m,b) \quad . \tag{10}$$

likelihood, probability, and objective functions

$$L(m,b|\vec{y}) = \prod_i^N p_i(y_i|x_i,m,b)$$

$$L(m,b|\vec{y}) = \prod_i^N p_i(y_i|x_i,\sigma_i,m,b)$$

$$p(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp - \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}$$

$$\log a \cdot b = \log a + \log b$$

$$\ln L(m,b|\vec{y}) = \ln \prod_i^N \frac{1}{\sigma_i\sqrt{2\pi}} \exp - \frac{y_i-(mx_i+b)}{2\sigma_i^2}$$

likelihood, probability, and objective functions

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, m, b)$$

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, \sigma_i, m, b)$$

$$p(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp - \frac{(y_i - (mx_i + b))^2}{2\sigma_i^2}$$

$$x^a \cdot x^b = x^{(a+b)}$$

$$\ln L(m, b|\vec{y}) = \ln \prod \frac{1}{\sigma_i\sqrt{2\pi}} + \ln\left(\prod_i^N e^{-\frac{y_i - (mx_i + b)}{2\sigma_i^2}}\right)$$

likelihood, probability, and objective functions

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, m, b)$$

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, \sigma_i, m, b)$$

$$p(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp - \frac{(y_i - (mx_i + b))^2}{2\sigma_i^2}$$

$$\ln L(m, b|\vec{y}) = \ln \prod \frac{1}{\sigma_i\sqrt{2\pi}} + \ln \left(e^{-\sum_i^N \frac{y_i - (mx_i + b)}{2\sigma_i^2}} \right)$$

likelihood, probability, and objective functions

$$L(m,b|\vec{y}) = \prod_i^N p_i(y_i|x_i,m,b)$$

$$L(m,b|\vec{y}) = \prod_i^N p_i(y_i|x_i,\sigma_i,m,b)$$

$$p(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp - \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}$$

σ_i not part of the model

$$\ln L(m,b|\vec{y}) = \ln \prod \frac{1}{\sigma_i\sqrt{2\pi}} + \ln\left(e^{-\sum_i^N \frac{y_i-(mx_i+b)}{2\sigma_i^2}}\right)$$

likelihood, probability, and objective functions

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, m, b)$$

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, \sigma_i, m, b)$$

$$p(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp -\frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}$$

$$\ln L(m, b|\vec{y}) = K - \sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}$$

likelihood, probability, and objective functions

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, m, b)$$

$$L(m, b|\vec{y}) = \prod_i^N p_i(y_i|x_i, \sigma_i, m, b)$$

$$p(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp - \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2}$$

$$\ln L(m, b|\vec{y}) = K - \sum \frac{(y_i-(mx_i+b))^2}{2\sigma_i^2} = K - \frac{1}{2}\chi^2$$

http://www-sop.inria.fr/members/Ian.Jermyn/philosophy/writings/Boxonmaths.pdf

its long,

but you have 2 weeks!

reading

Science and Statistics

GEORGE E. P. BOX*

Aspects of scientific method are discussed: In particular, its representation as a motivated iteration in which, in succession, practice confronts theory, and theory, practice. Rapid progress requires sufficient flexibility to profit from such confrontations, and the ability to devise parsimonious but effective models, to worry selectively about model inadequacies and to employ mathematics skillfully but appropriately. The development of statistical methods at Rothamsted Experimental Station by Sir Ronald Fisher is used to illustrate these themes.

1. INTRODUCTION

In 1952, when presenting R.A. Fisher for the Honorary degree of Doctor of Science at the University of Chicago, W. Allen Wallis described him in these words.

> He has made contributions to many areas of science; among them are agronomy, anthropology, astronomy, bacteriology, botany, economics, forestry, meteorology, psychology, public health, and—above all—genetics, in which he is recognized as one of the leaders. Out of this varied scientific research and his skill in mathematics, he has evolved systematic principles for

on the one hand, nor by the undirected accumulation of practical facts on the other, but rather by a motivated *iteration* between theory and practice such as is illustrated in Figure A (1).

A. The Advancement of Learning

A(1) An Iteration Between Theory and Practice

A(2) A Feedback Loop