Urban Informatics

Fall 2015

dr. federica bianco fb55@nyu.edu

@fedhere

Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing: *p*-value, statistical significance
- Statistical and Systematic errors
- Goodness of fit tests

Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing: *p*-value, statistical significance
- Statistical and Systematic errors
- Goodness of fit tests

Today:

- Residuals minimization
- Likelihood
- model diagnostics
  Chi^2, R^2, and LR test
- Higher degree regression

Goodness of fit

You have some data, and an idea of how it should look: a *model*

Is it a good model?

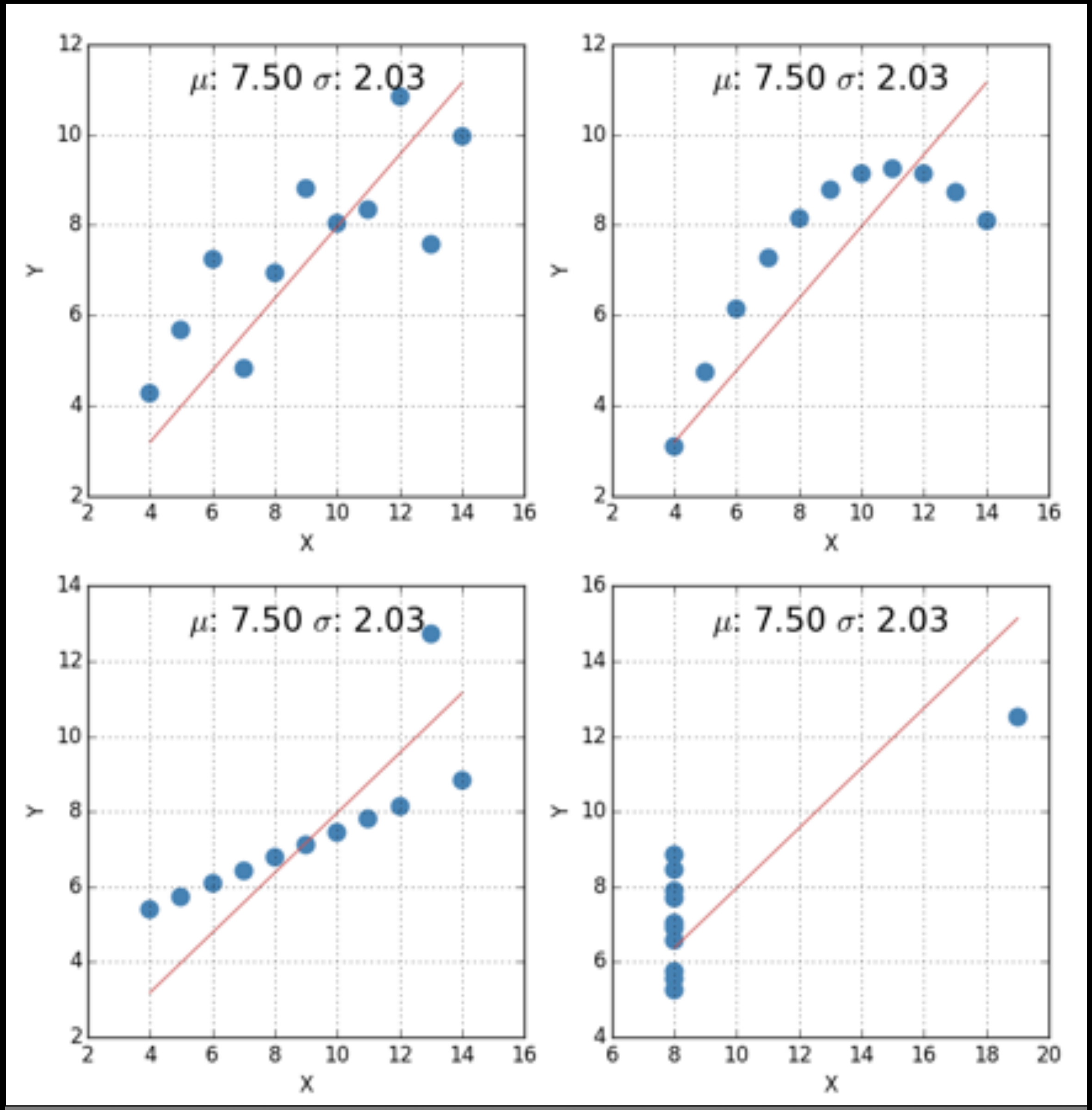
Goodness of fit

Tests Cheat Sheet:
goodness of fit

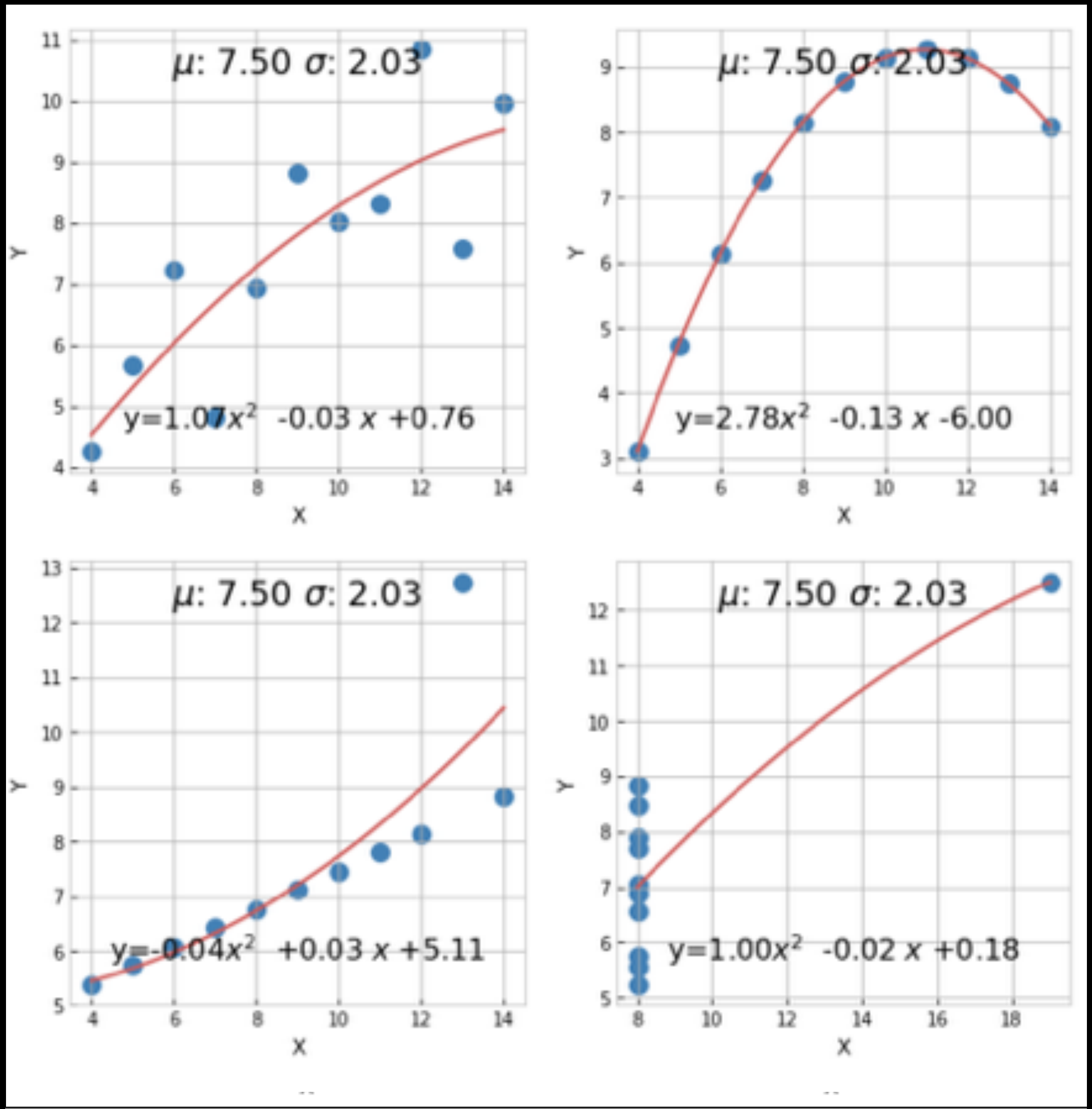
	metric (statistic)	compare to	
KS	$D_{n_1,n_2}(x) = max(\mid F_n(x) - F\ (x) \mid)$	$\frac{K_\alpha}{\sqrt{n}}$	power in the core only
Pearson's chi square	$\chi^2_{red} = \frac{\chi^2}{df} = \frac{1}{df} \sum \frac{(O-E)^2}{\sigma^2}$	scipy.stats.chisquare(f_obs, f_exp=None, ddof=0, axis=0)[	
Anderson-Darling	$A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)\ (1-F(x))} dF(x)$	scipy.stats.anderson(x, dist='norm')	power in the tails
K-L divergence	$D_{KL} = - \int_x p(x)\ log(q(x)) + p(x)\ log(p(x))$	scipy.stats.entropy(pk. qk=<not None>)	relates to information entropy
Likelihood ratio	$\frac{L\ (model\ 1 \mid\ data)}{L\ (model\ 2 \mid\ data)}$		suitable to bayesian analysis

What's a model??

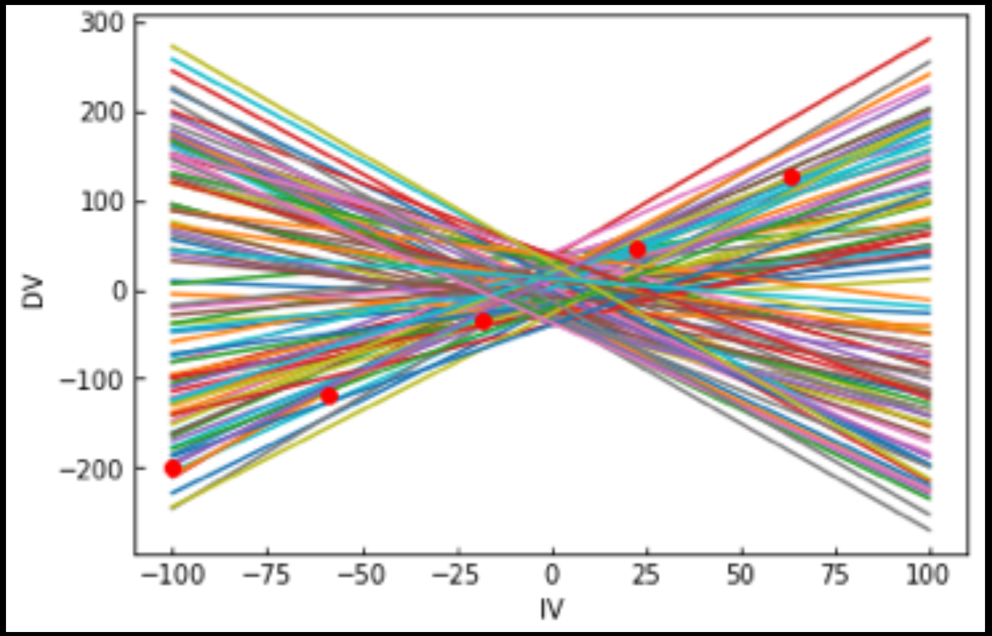
What's a model??



What's a model??



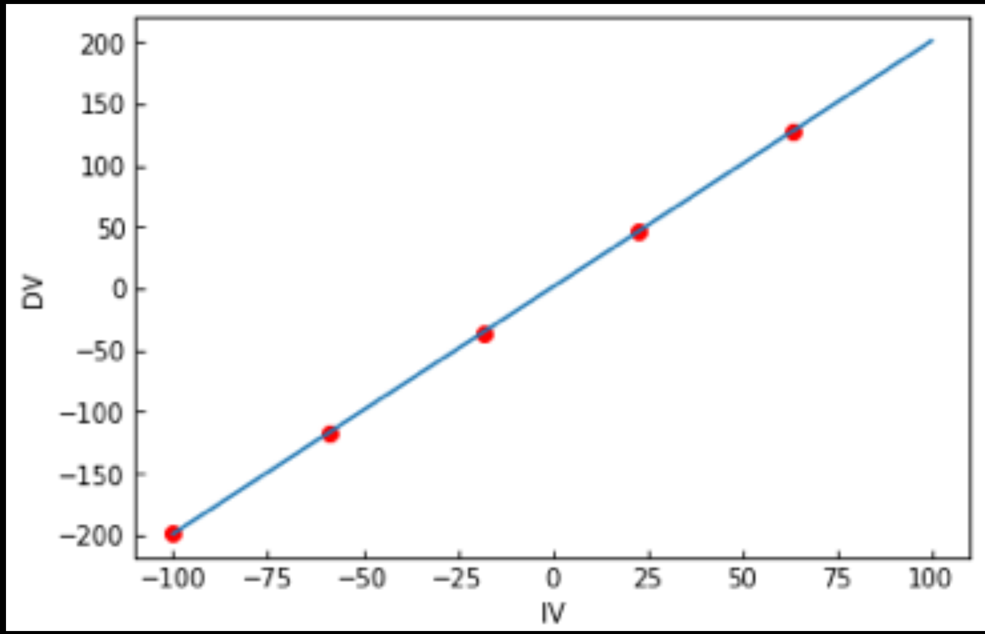
a formula that describes the data ⟶ *a family of models*



What's a model??

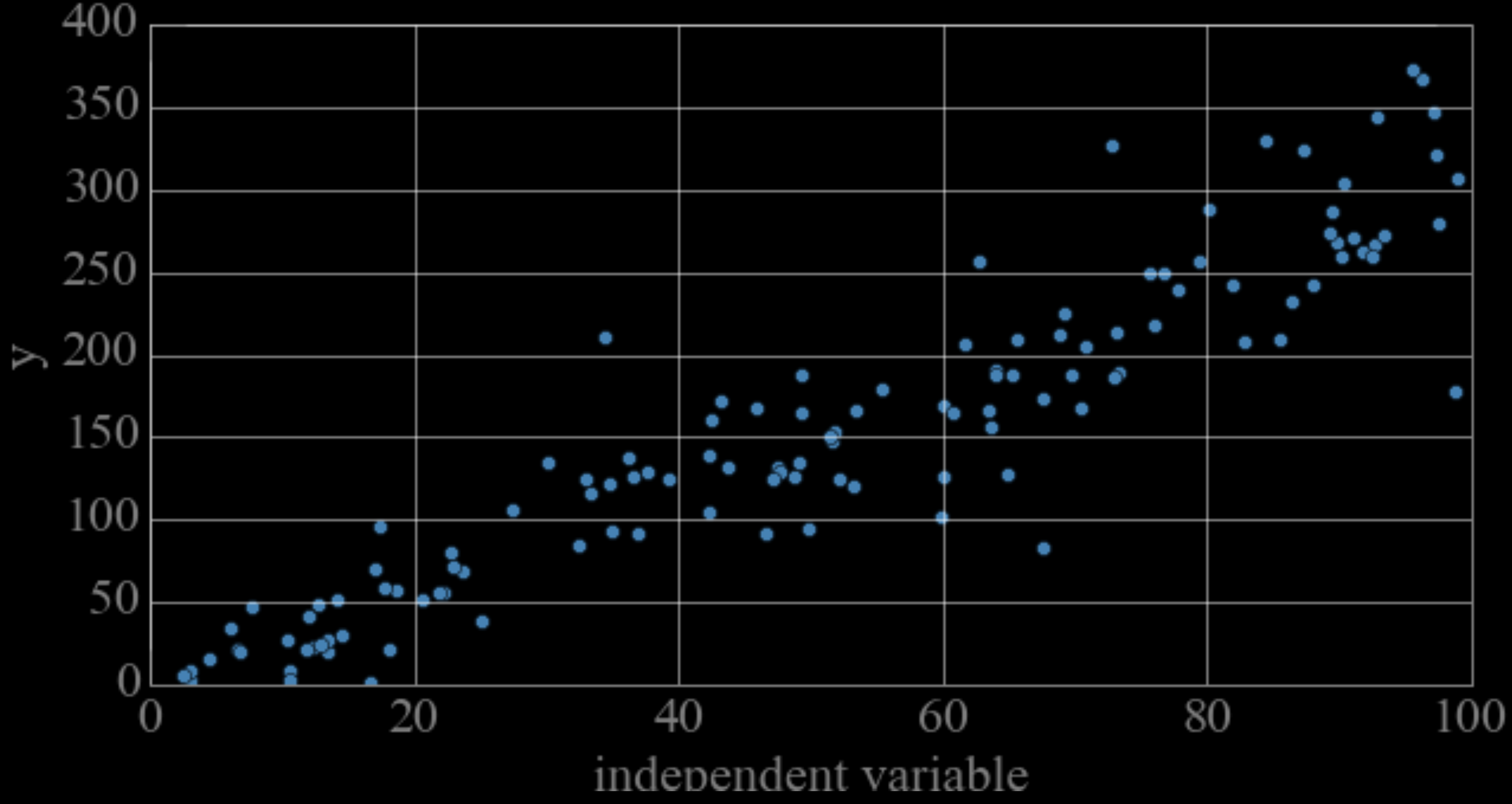
a formula that describes the data ⟶ *a family of models*

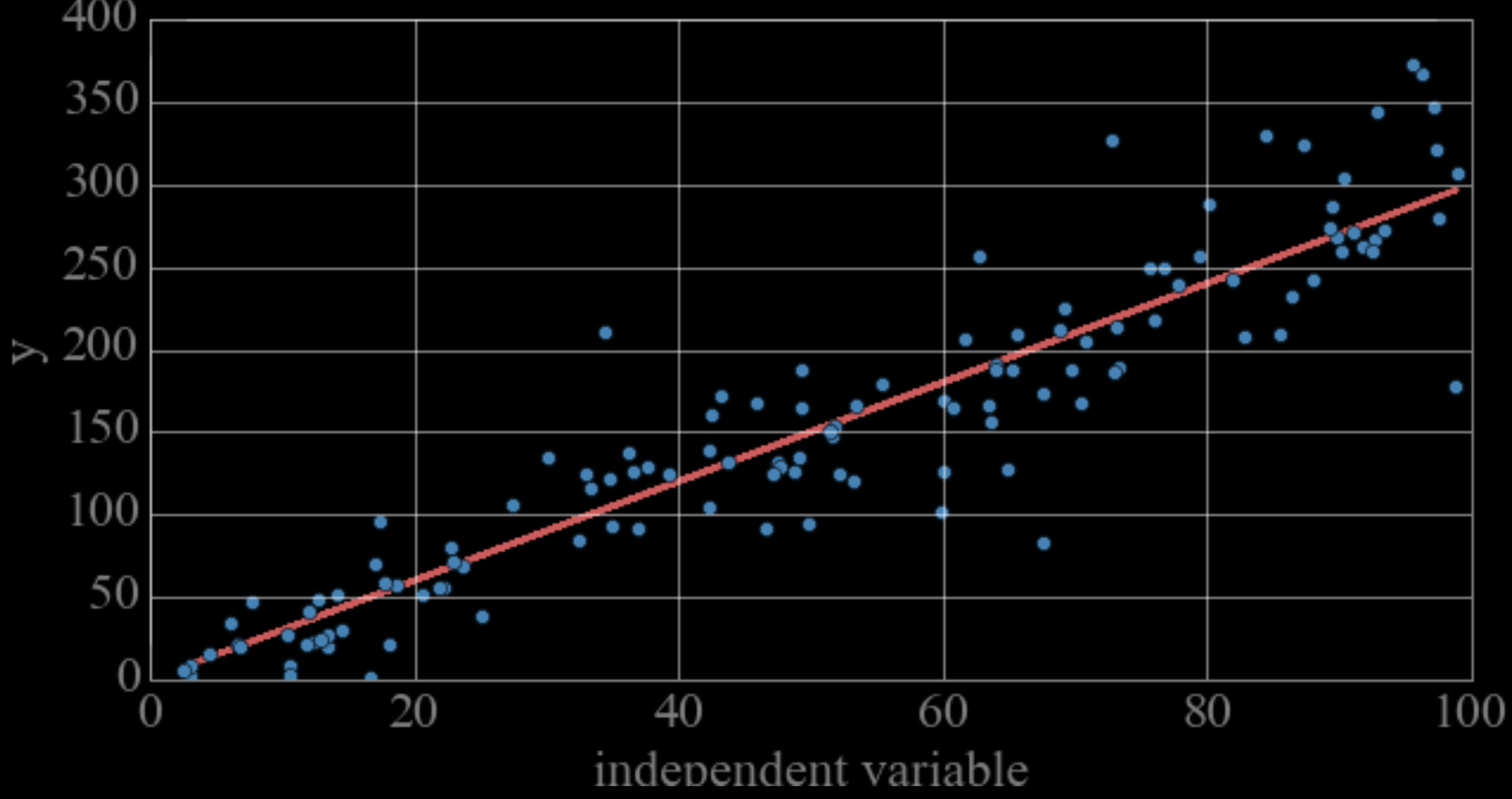
the best fit chooses within that family
the model that has the
best parameters



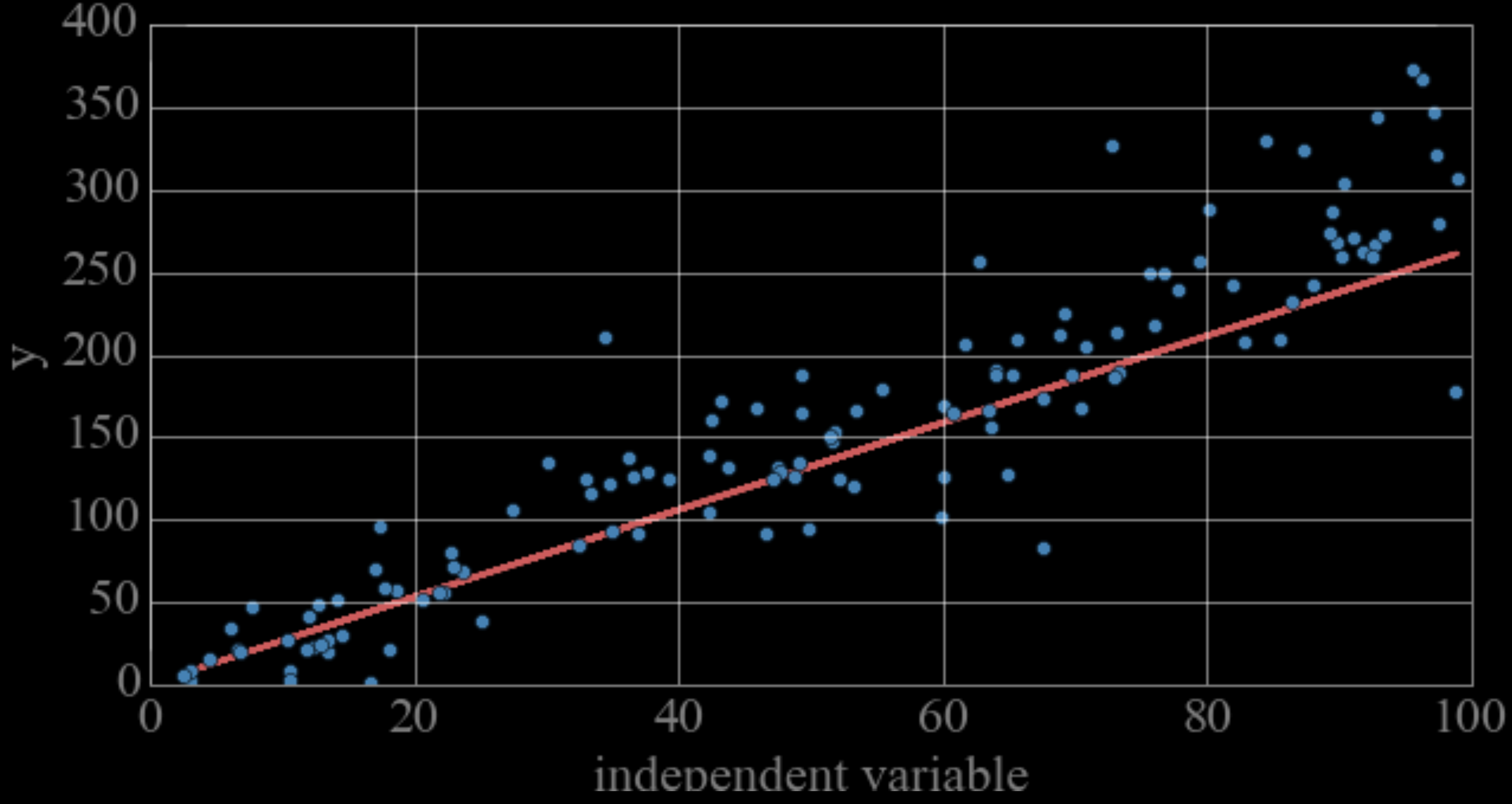
What's a model??

How do we fit a model to data?

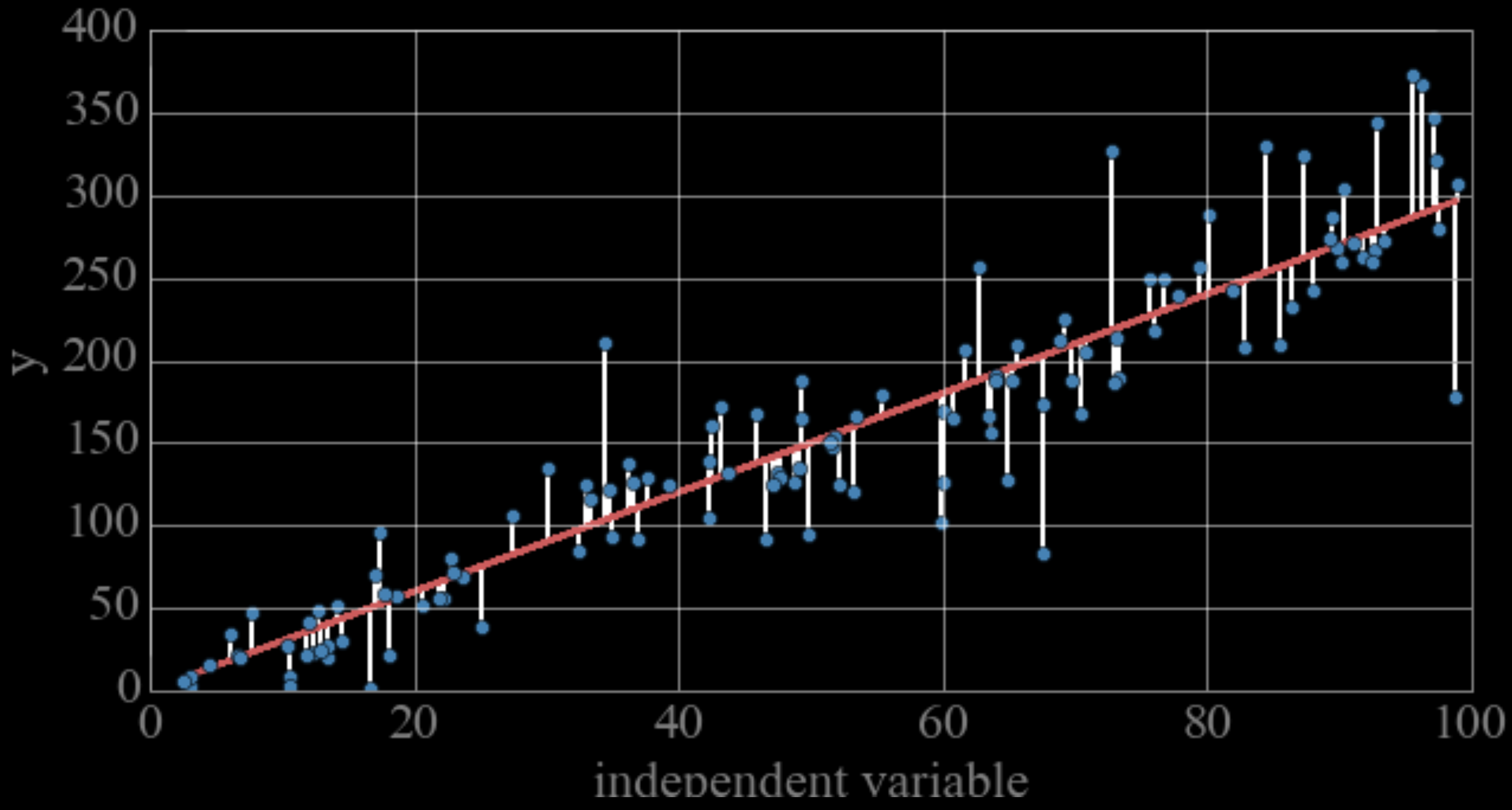

400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable

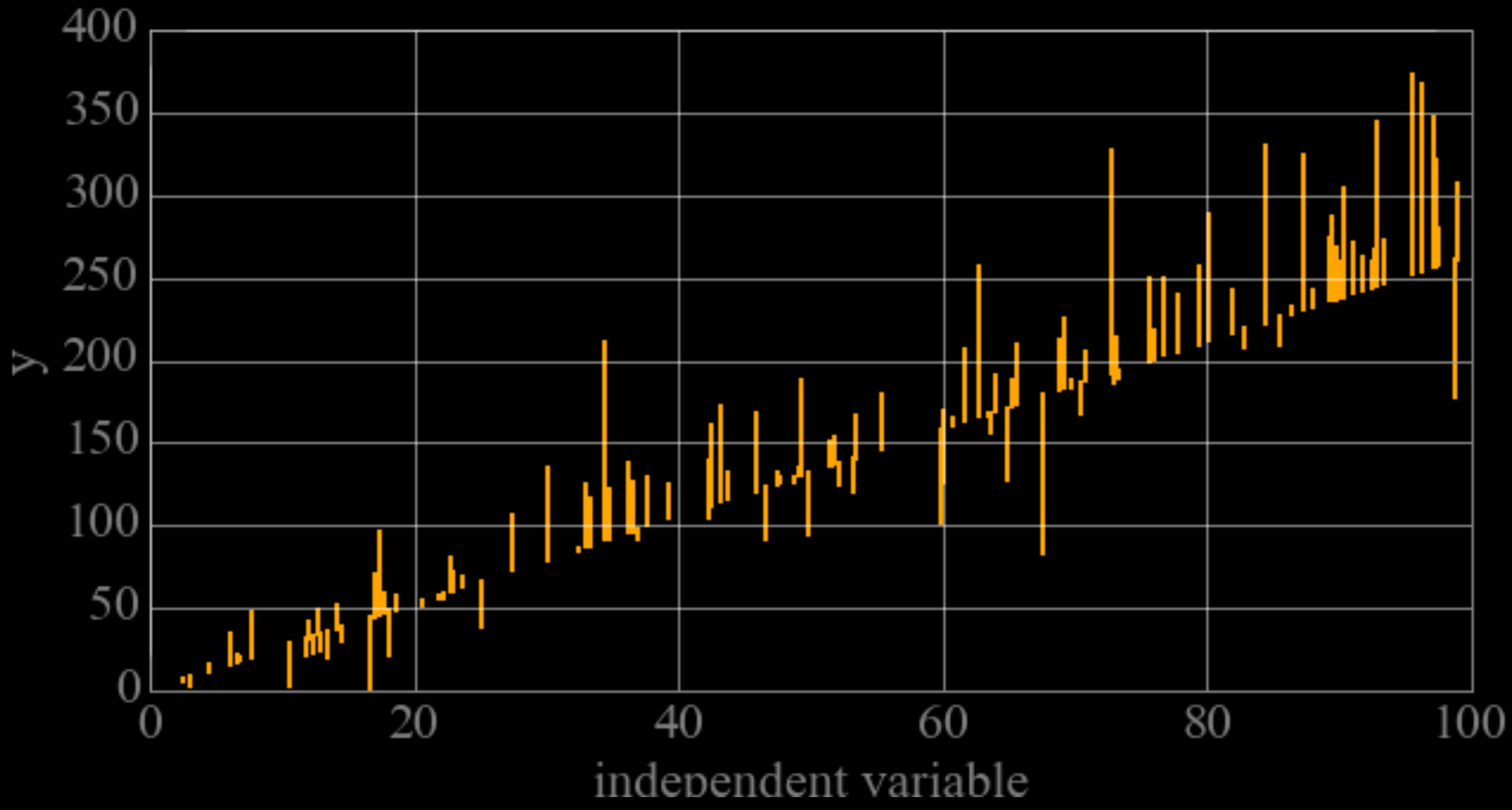


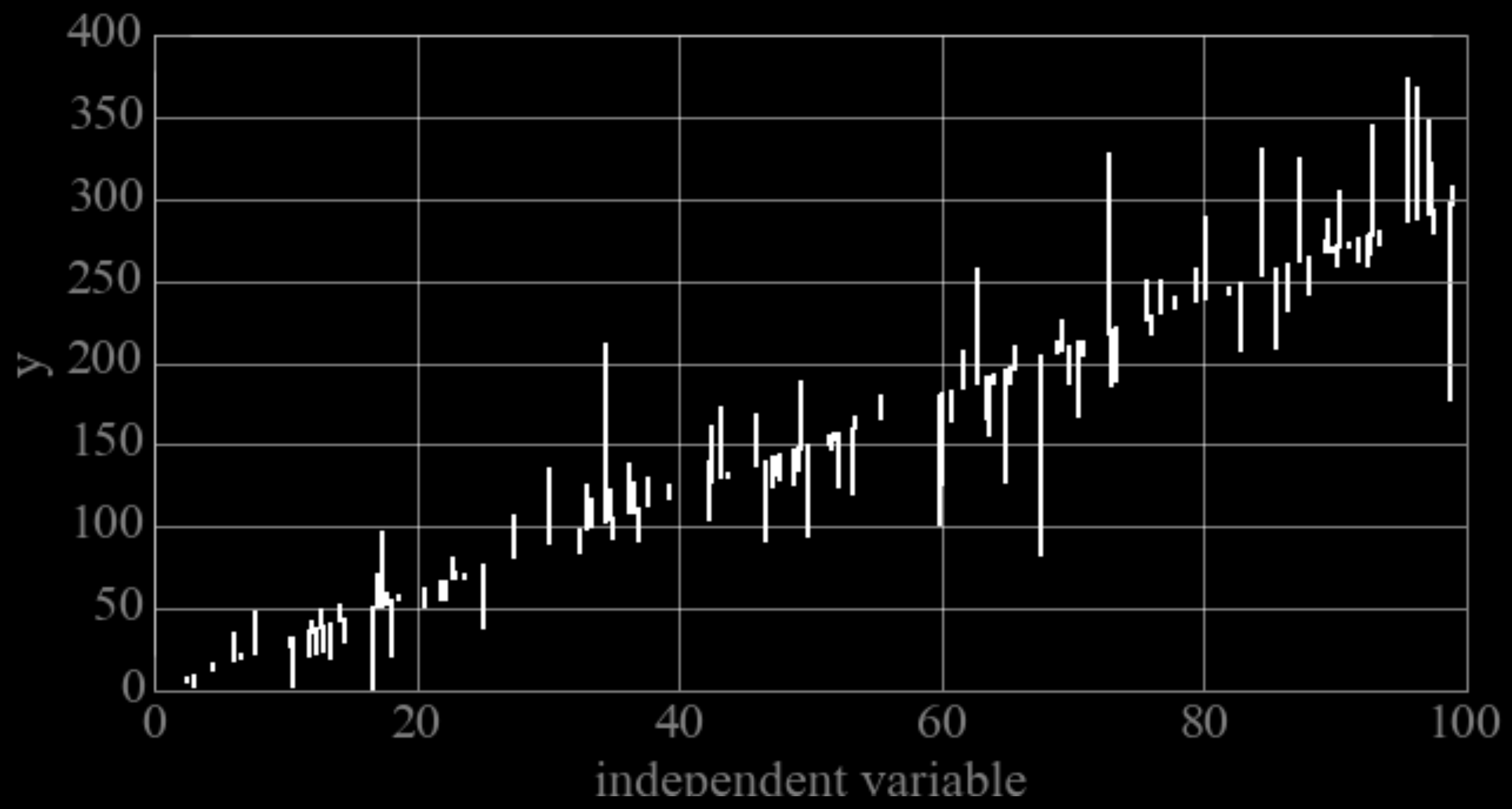
400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable



400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable

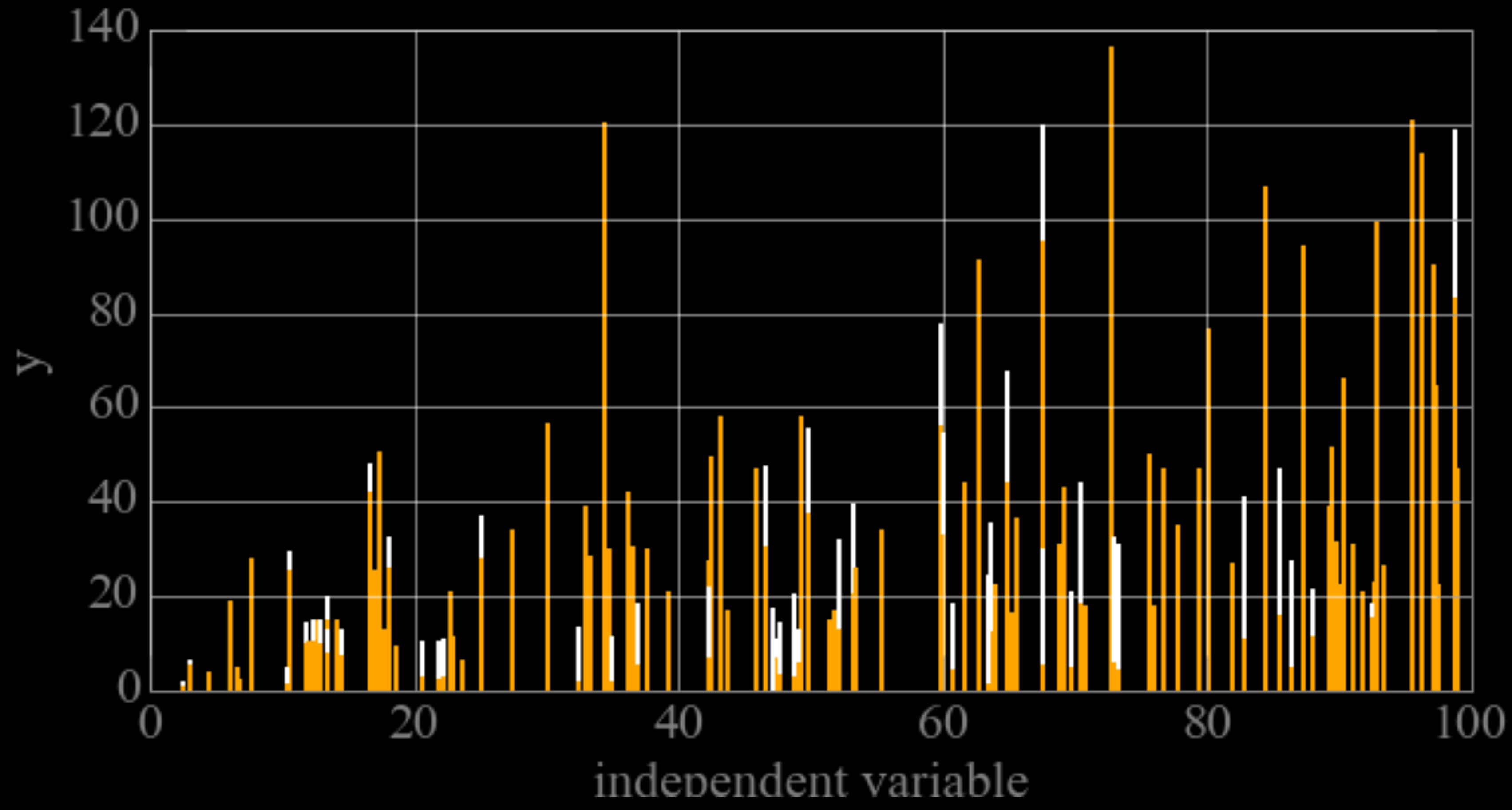

400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable


400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable


400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable

Fit model parameters =
minimize the
Sum of residuals squared $\sum_i (y_i - (ax_i + b))^2$

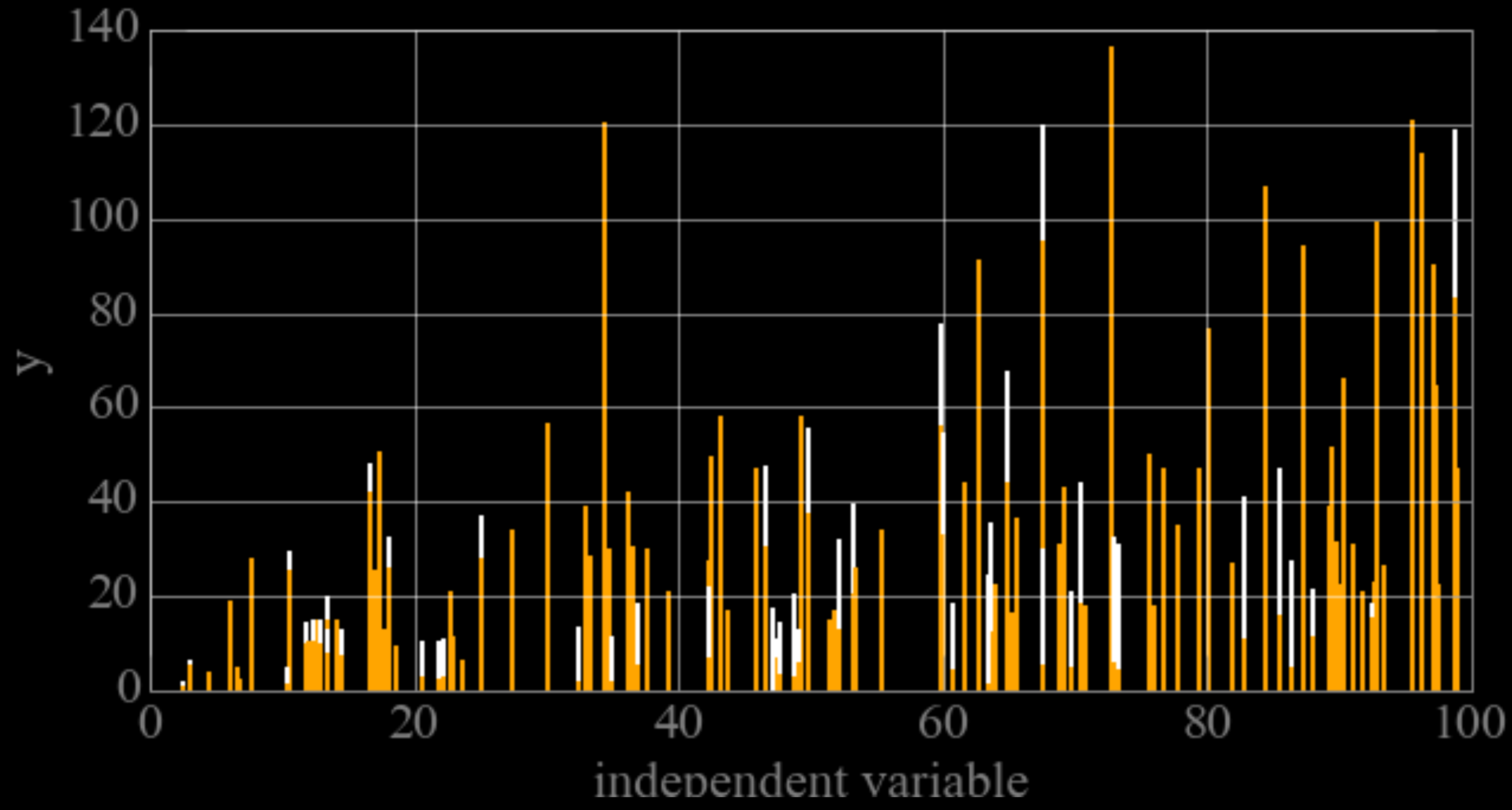
11655.34 < 12155.24



Fit model parameters =

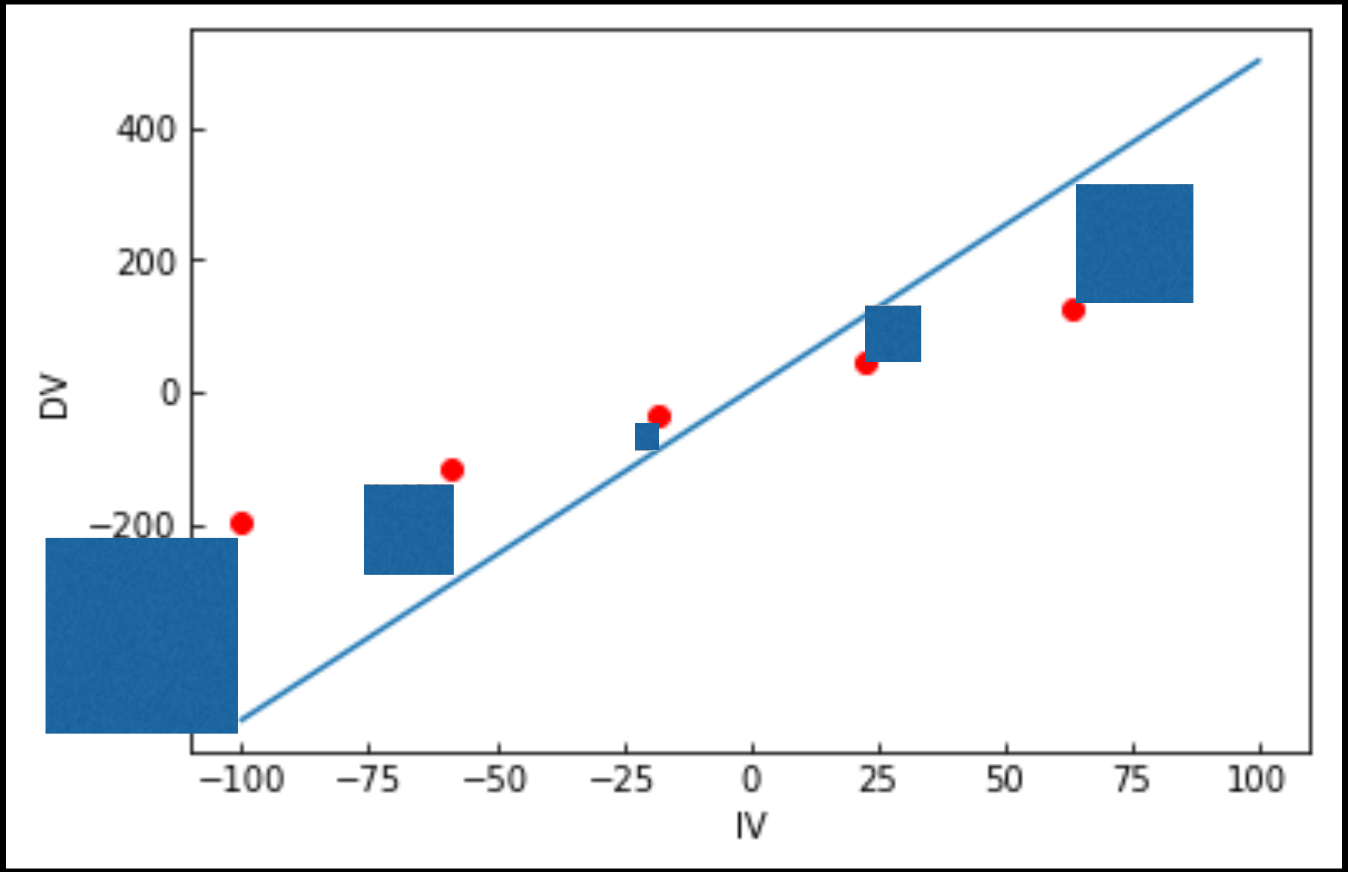
find *m* and *b such that* $\sum_i (y_i - (ax_i + b))^2$ *is minimal*

11655.34 < 12155.24



R^2 $\sum_i (y_i - (ax_i + b))^2$

amount of variance in data explained by the model



OLS Regression Results

Dep. Variable:	Y	R-squared:	0.687
Model:	OLS	Adj. R-squared:	0.609
Method:	Least Squares	F-statistic:	8.793
Date:	Tue, 11 Oct 2016	Prob (F-statistic):	0.00956
Time:	06:14:52	Log-Likelihood:	-16.487
No. Observations:	11	AIC:	38.97
Df Residuals:	8	BIC:	40.17
Df Model:	2		
Covariance Type:	nonrobust		

adjusted R²

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

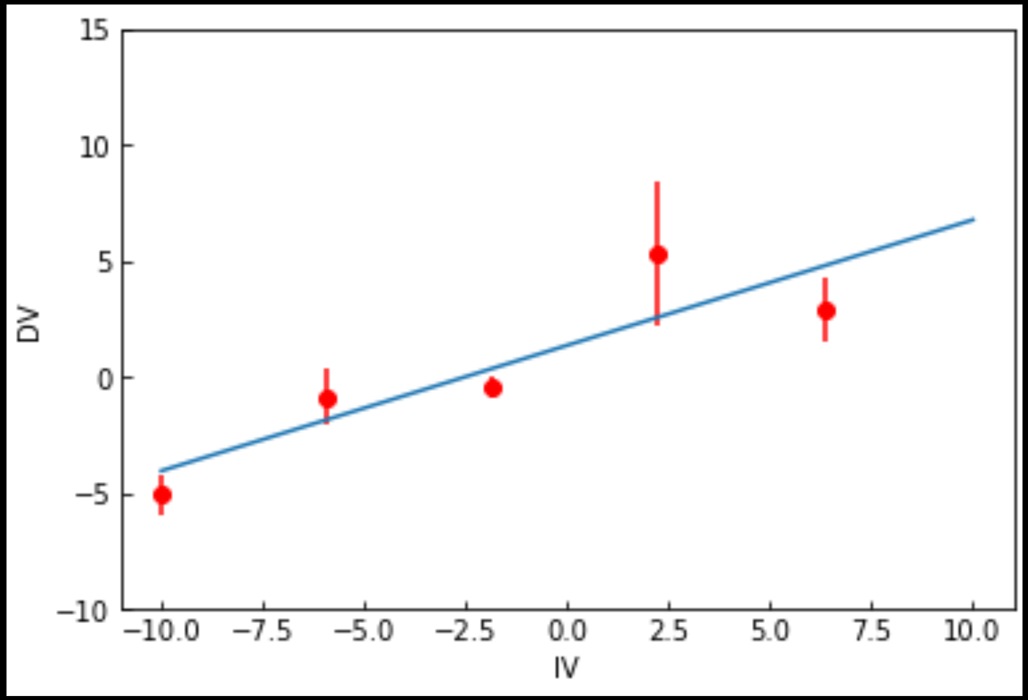
adjusts for the number of *explanatory terms* (parameters) in a model relative to the number of data points

χ^2 (chi²)

$$\chi^2_{/DOF} = \frac{1}{DOF} \sum_i \frac{(m_i - x_i)^2}{\sigma_i^2}$$

$R^2 = 0.8$

$\chi^2 = 2.5$



χ^2 (chi^2)

$$\chi_F^2 = \sum_i \frac{(m_i - x_i)^2}{\sigma_i^2}$$

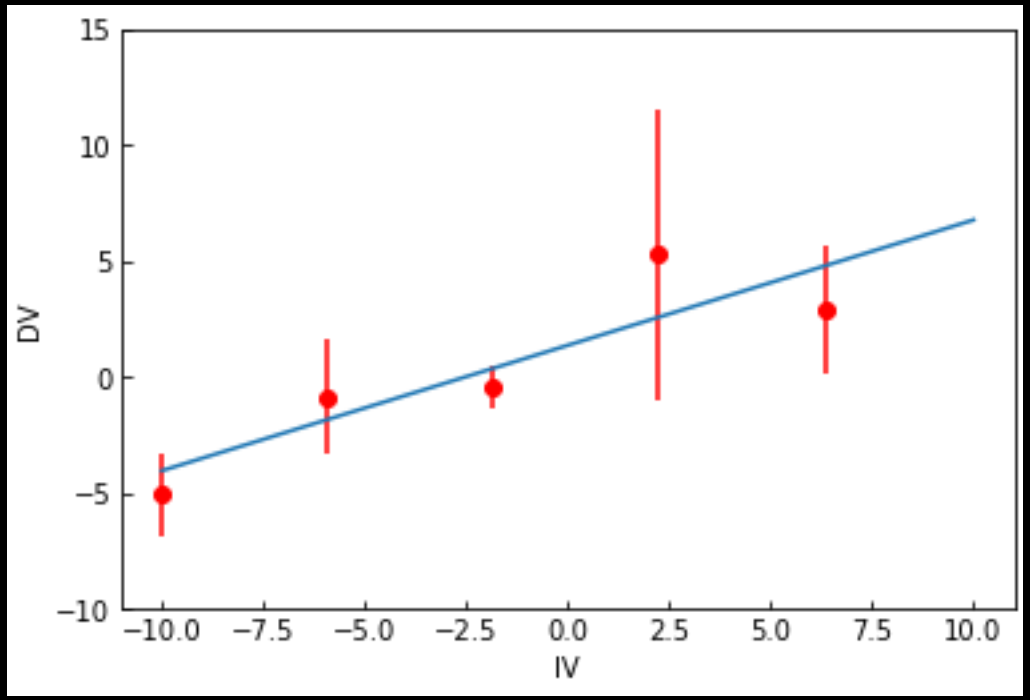
how well model explains data *including uncertainties*

χ^2 (chi²)

$$\chi^2_{/DOF} = \frac{1}{DOF} \sum_i \frac{(m_i - x_i)^2}{\sigma_i^2}$$

$R^2 = 0.8$

$\chi^2 = 0.6$

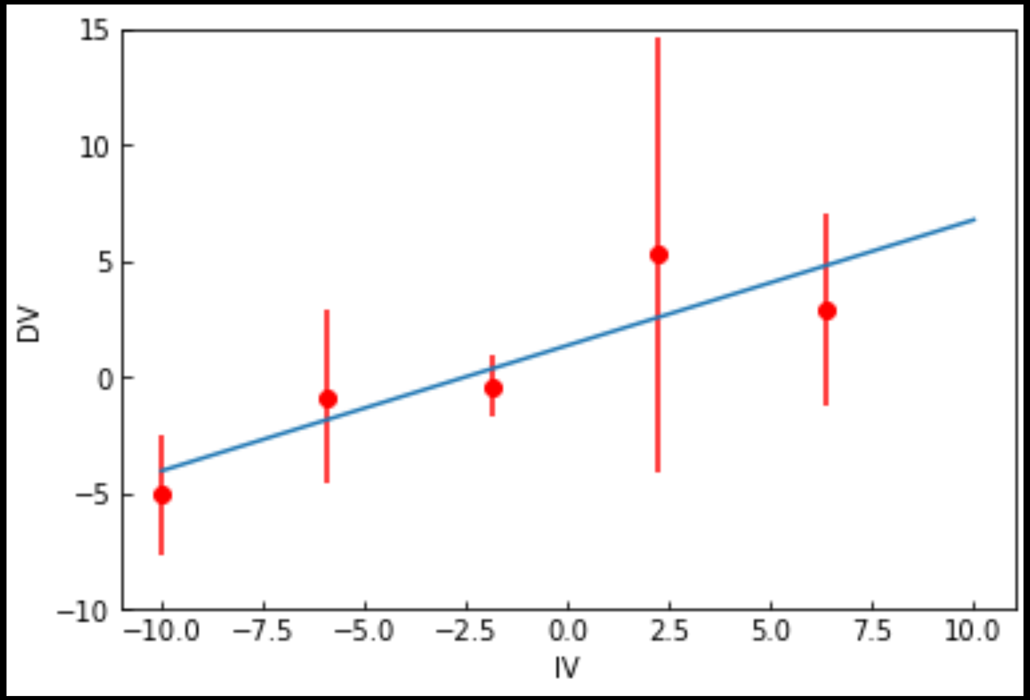


χ^2 (chi²)

$$\chi^2_{/DOF} = \frac{1}{DOF} \sum_i \frac{(m_i - x_i)^2}{\sigma_i^2}$$

$R^2 = 0.8$

$\chi^2 = 0.3$



Likelihood

Probability $P(\; x \mid \theta \;)$

Likelihood

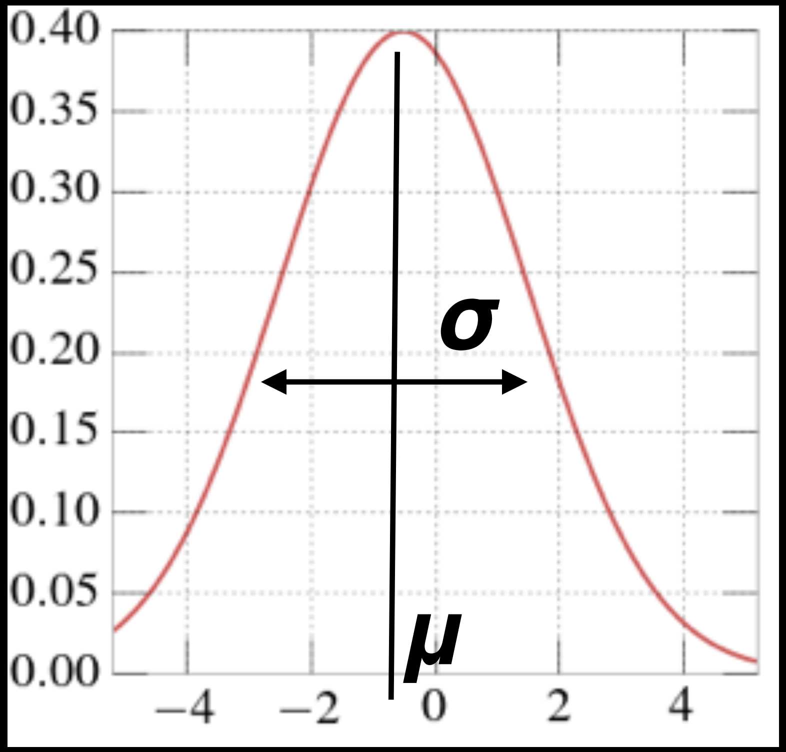
Probability $P(\vec{x} \mid \mu, \sigma)$

Likelihood

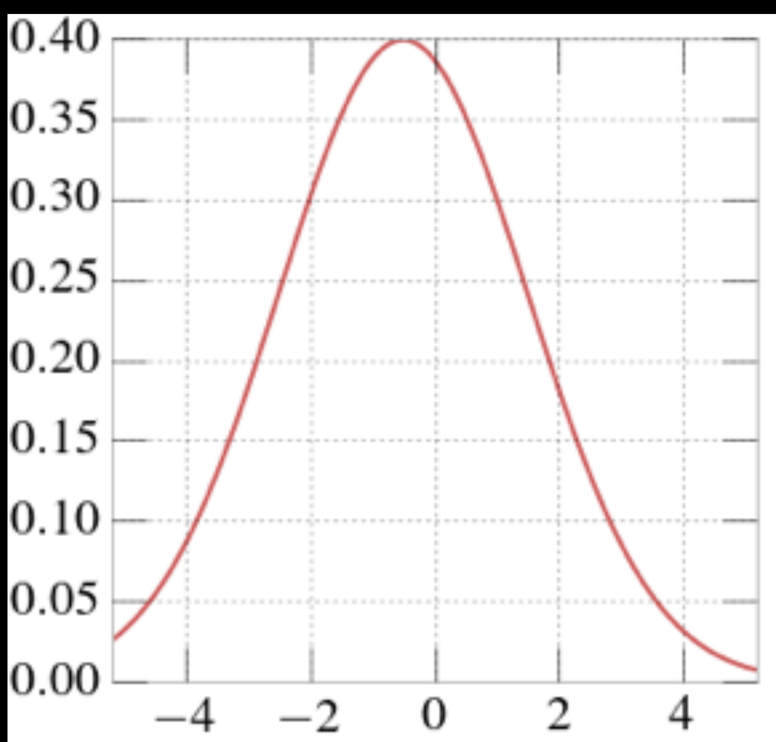
Probability $P(\vec{y} \mid \vec{x}, \mu, \sigma)$

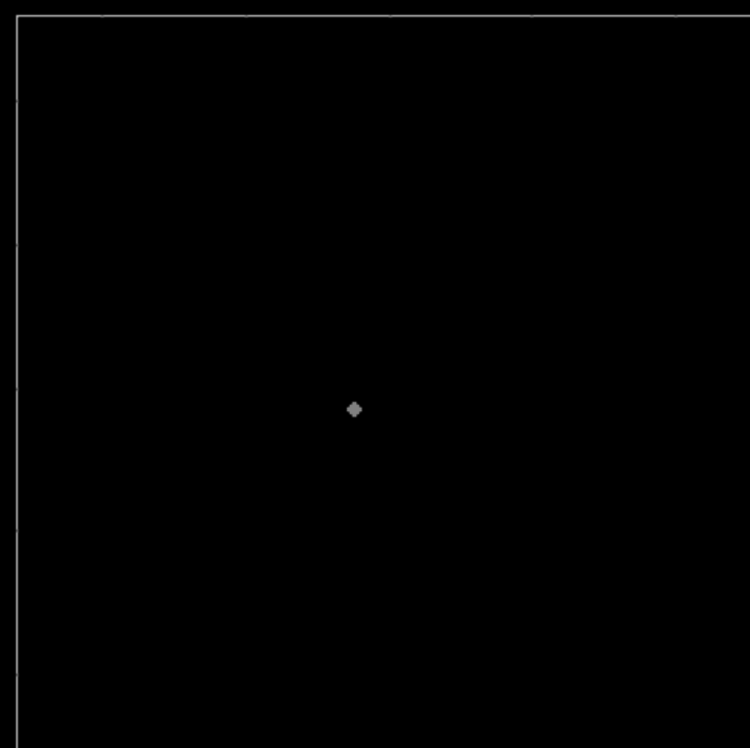
Likelihood

Probability $P(\vec{x} \mid \mu, \sigma)$

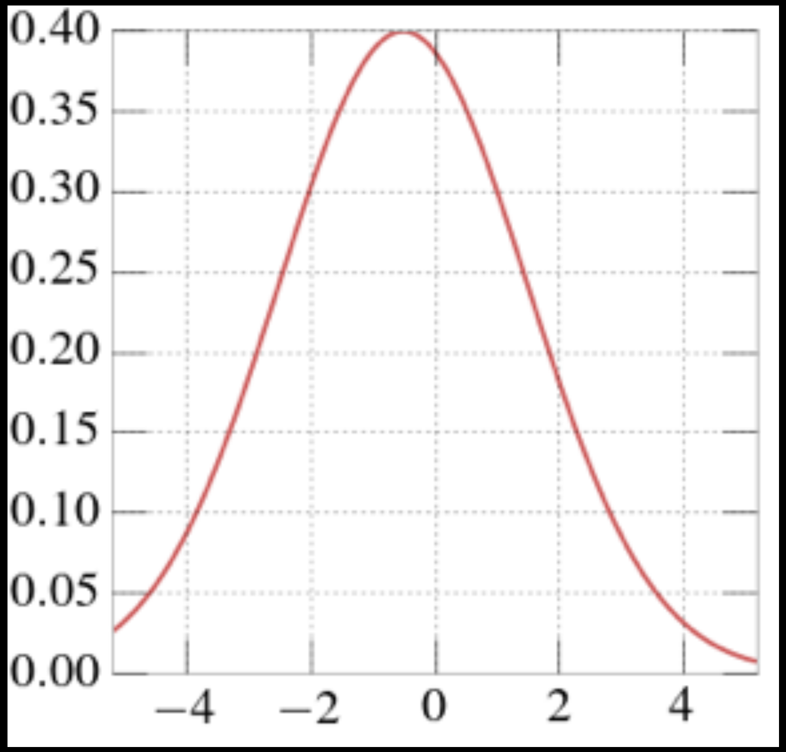


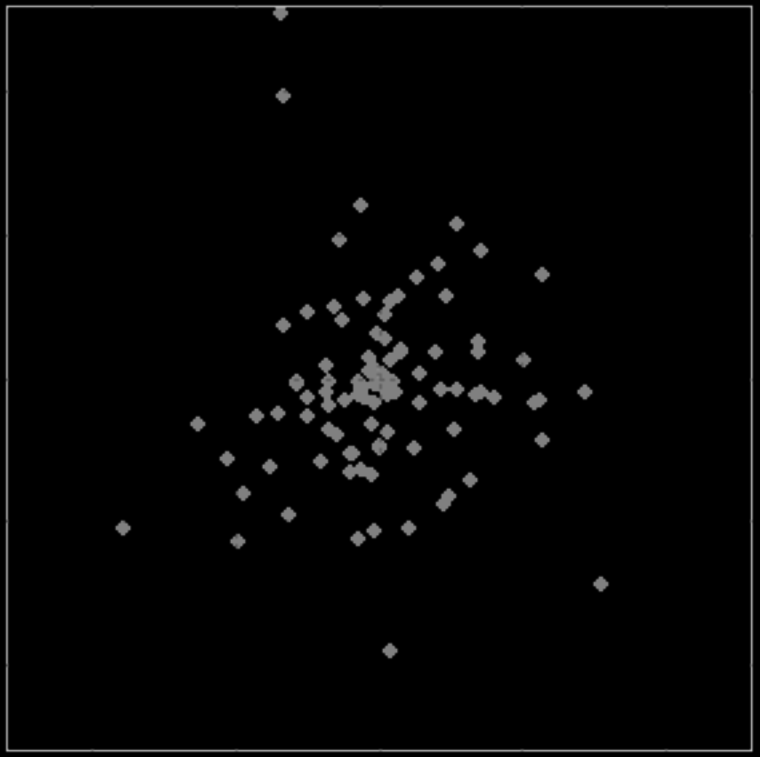
Probability $P(\vec{x} \mid \vec{\theta})$

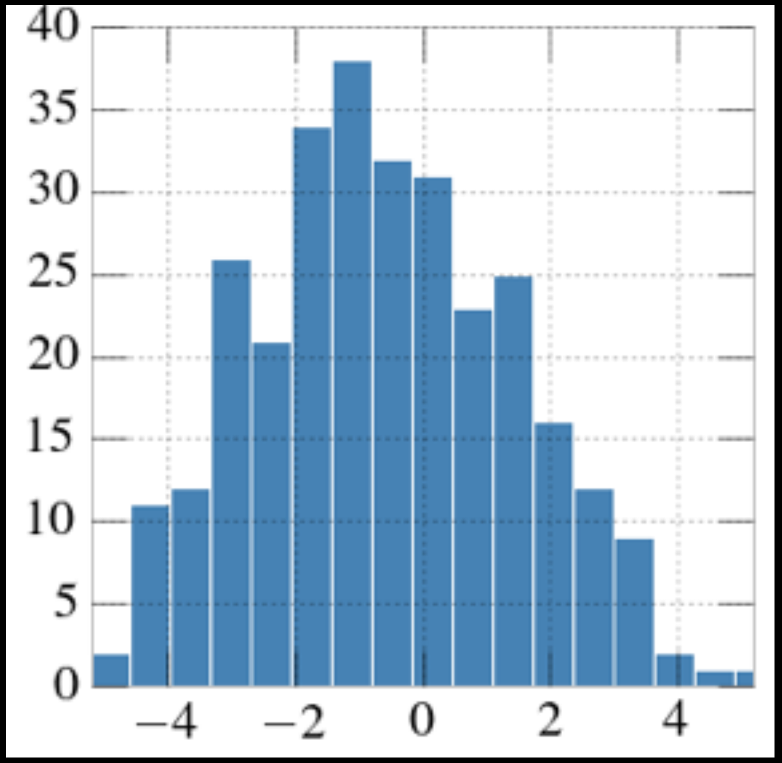




Probability $P(\vec{x} \mid \vec{\theta})$

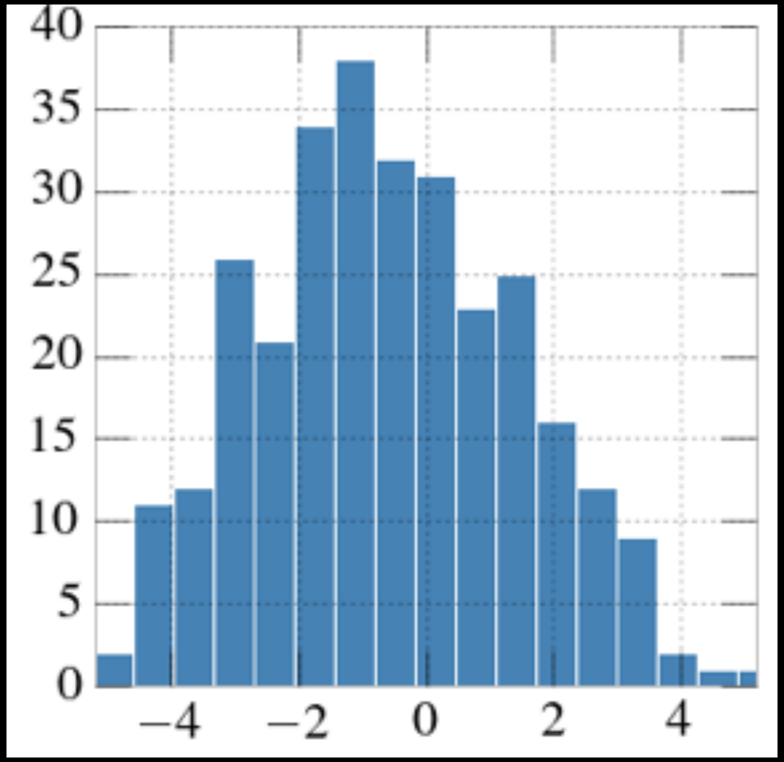






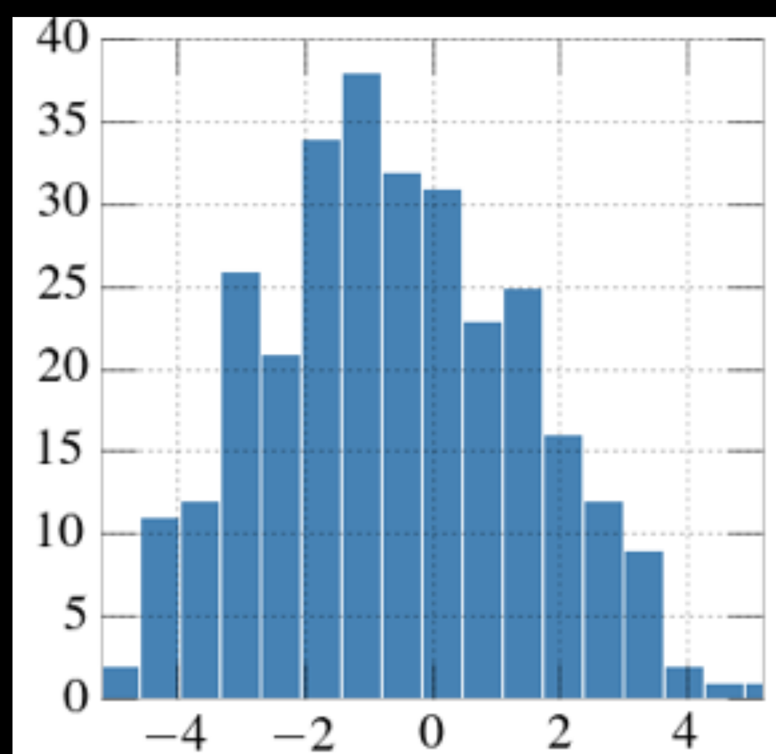
Probability $P(\vec{x} \mid \vec{\theta})$

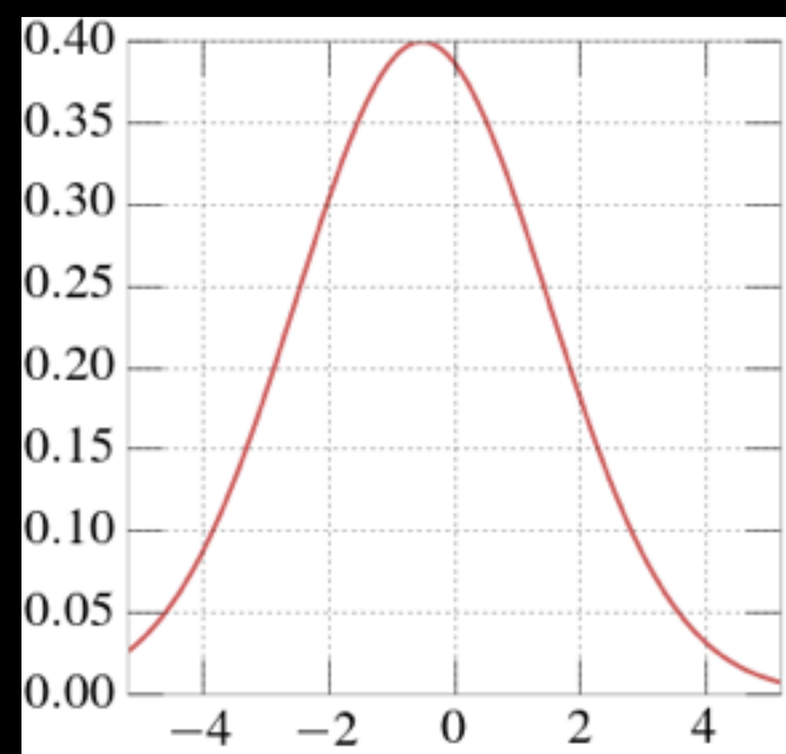
Likelihood $P(\vec{\theta} \mid \vec{x})$



Probability $P(\vec{x} \mid \vec{\theta})$

Likelihood $P(\vec{\theta} \mid \vec{x})$





Probability $$N(\mu,\sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Likelihood

Probability $N(\mu,\sigma) \sim \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Likelihood $\mathscr{L}_{(\mu,\sigma)}(x) \sim \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

CUSP

Probability $$N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Likelihood

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Likelihood $\mathscr{L}_{(\mu,\sigma)}(\vec{x}) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Likelihood $\mathscr{L}_{(\mu,\sigma)}(\vec{x}) \sim \frac{1}{(\sigma\sqrt{2\pi})^n} \Pi_i e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Likelihood $\mathscr{L}_{(\mu,\sigma)}(\vec{x}) \sim \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\Sigma_i (x_i-\mu)^2}{2\sigma^2}}$

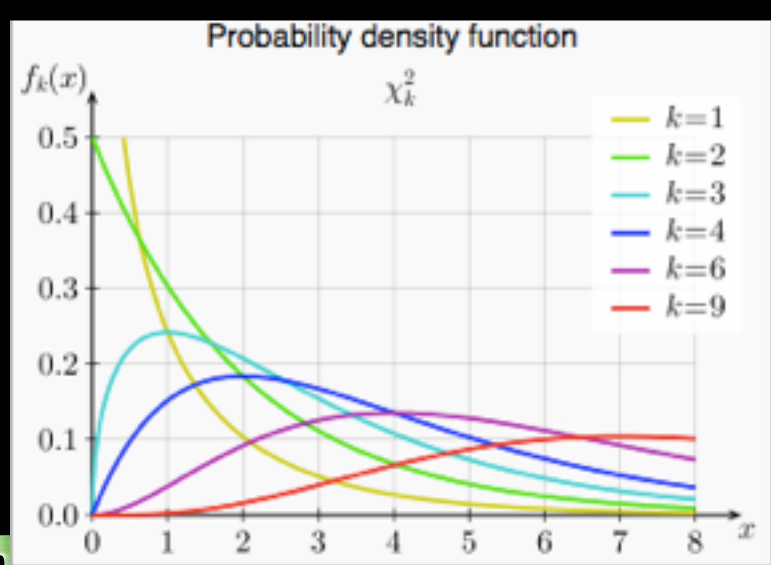
Likelihood-ratio tests

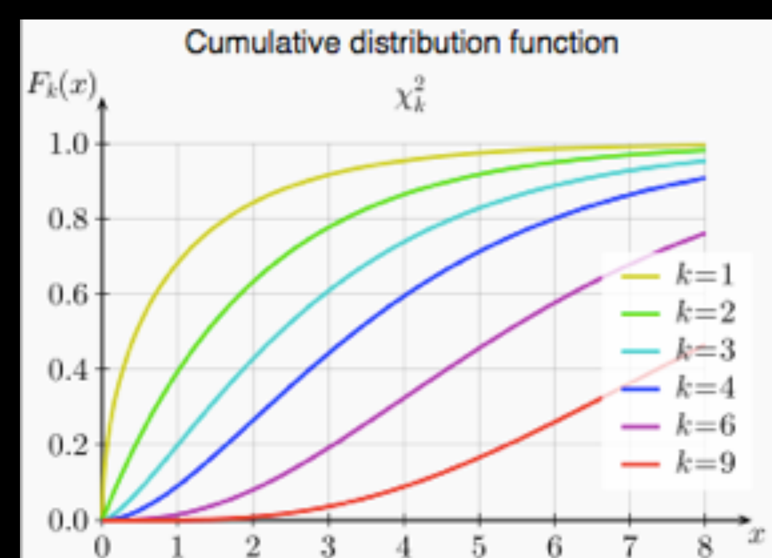
	H_0 is True	H_0 is False
H_0 is falsified	Type I error False Positive important message gets spammed	True Positive
H_0 is not falsified	True Negative	Type II error False negative Spam in your Inbox

$$LR = \frac{\text{False Negative}}{\text{True Negative}}$$

$$LR = -2 \log_e \frac{L(\text{model 1})}{L(\text{model 2})}$$

This statistic is <u>chi-squared distributed</u>





$$LR = -2 \log_e \frac{L(\text{model } 1)}{L(\text{model } 2)}$$

This statistic is <u>chi-squared distributed</u> with <u>degrees of freedom equal</u> to the <u>difference in the number of degrees of freedom between the two models</u> (i.e., the number of variables added to the model).

Maximizing Likelihood

Probability $$N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Likelihood $$\mathscr{L}_{(\mu,\sigma)}(\vec{x}) \sim \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\Sigma_i(x_i-\mu)^2}{2\sigma^2}}$$

Given some observations $\vec{x}$ we want to model them with the best function: the one that is MAXIMALLY LIKELY.

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Likelihood $\mathcal{L}_{(\mu,\sigma)}(\vec{x}) \sim \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\Sigma_i(x_i-\mu)^2}{2\sigma^2}}$

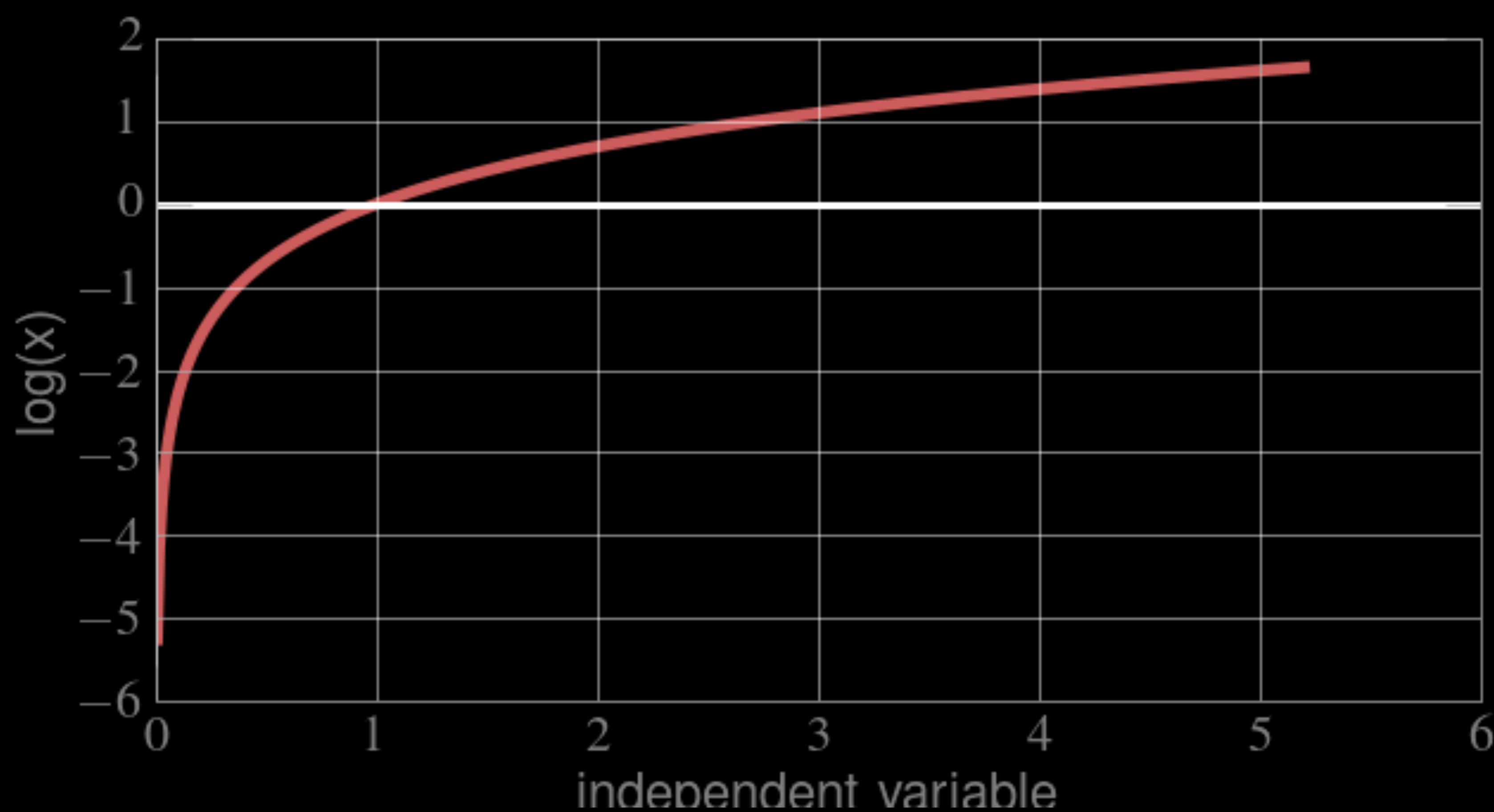
Given some observations $\vec{x}$ we want to model them with the best function: the one that is MAXIMALLY LIKELY. After we choose a functional form (N) for the model we want to choose the parameters (μ,σ) that maximiz $\mathcal{L}_{(\mu,\sigma)}(\vec{x})$

Probability $\quad N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

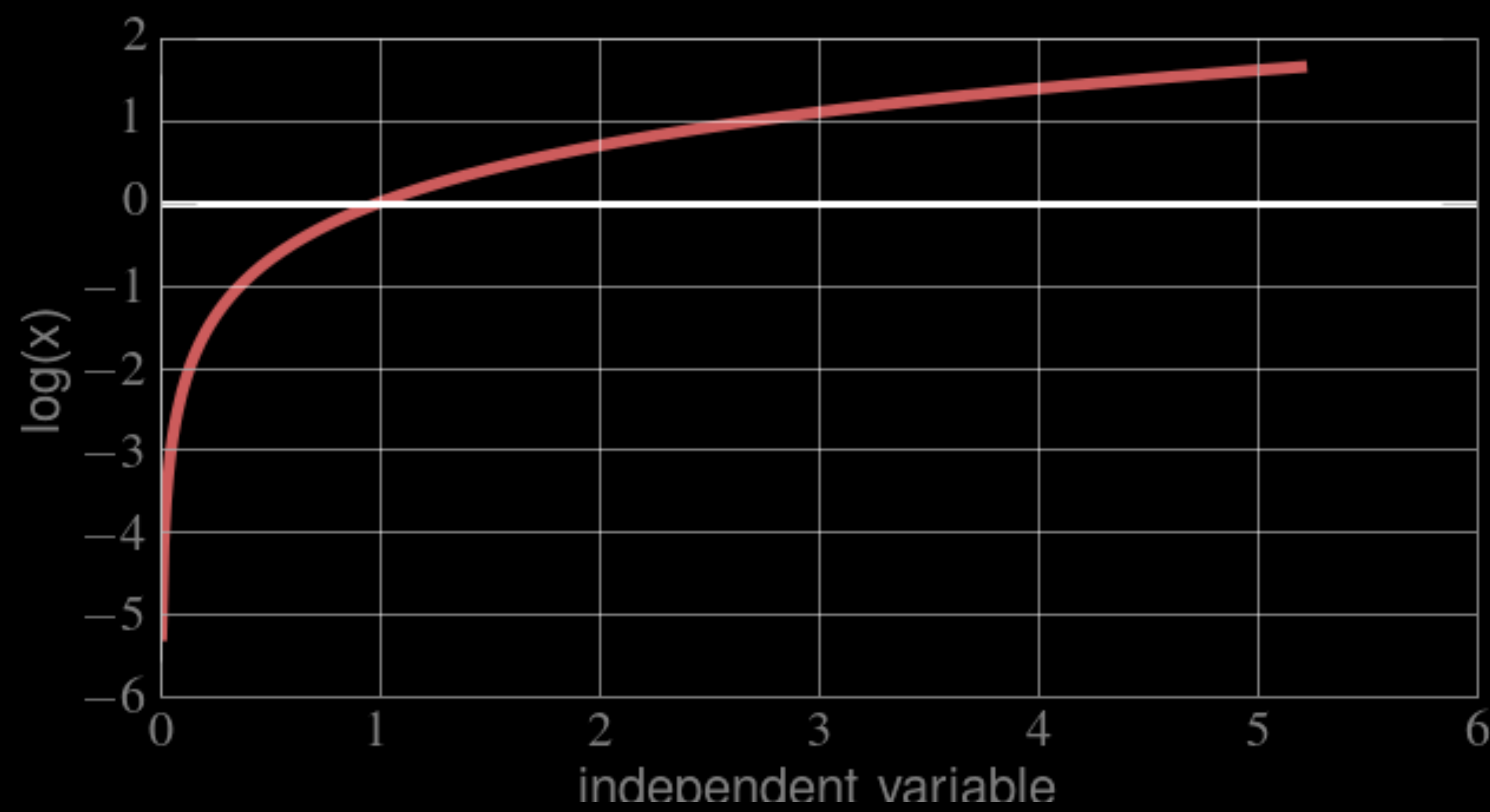
Likelihood $\quad \mathscr{L}_{(\mu,\sigma)}(\vec{x}) \sim \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\Sigma_i(x_i-\mu)^2}{2\sigma^2}}$

FIND $\mu^*, \sigma^* \mid \mathscr{L}_{(\mu^*,\sigma^*)} = \max(\mathscr{L}_{(\mu,\sigma)}(\vec{x}))$

Logarithm:



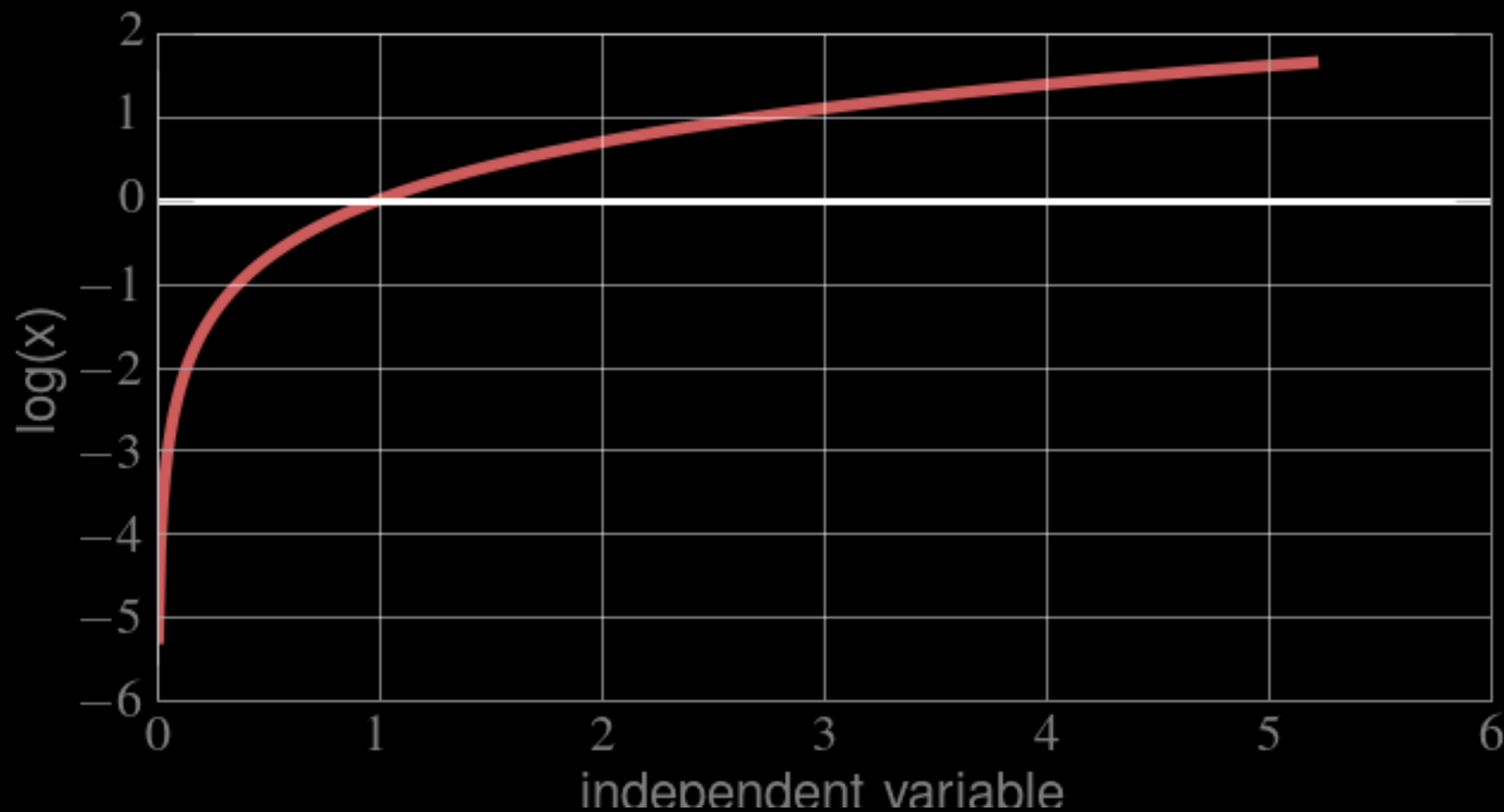
Logarithm: MONOTONICALLY INCREASING



Logarithm: MONOTONICALLY INCREASING

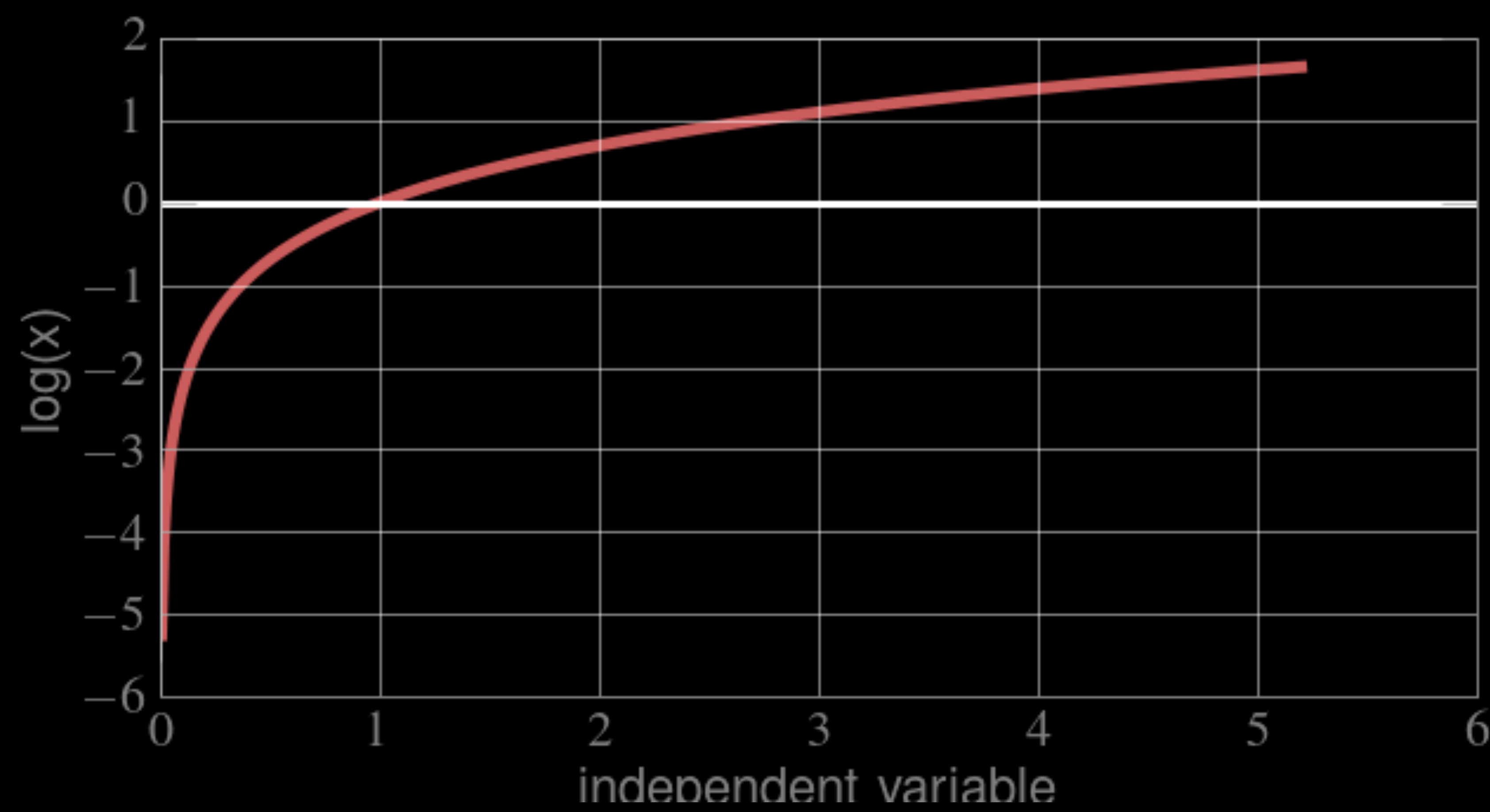
if x grows, log(x) grows, if x decreases, log(x) decreases

the location of the maximum is the same!



Logarithm: MONOTONICALLY INCREASING

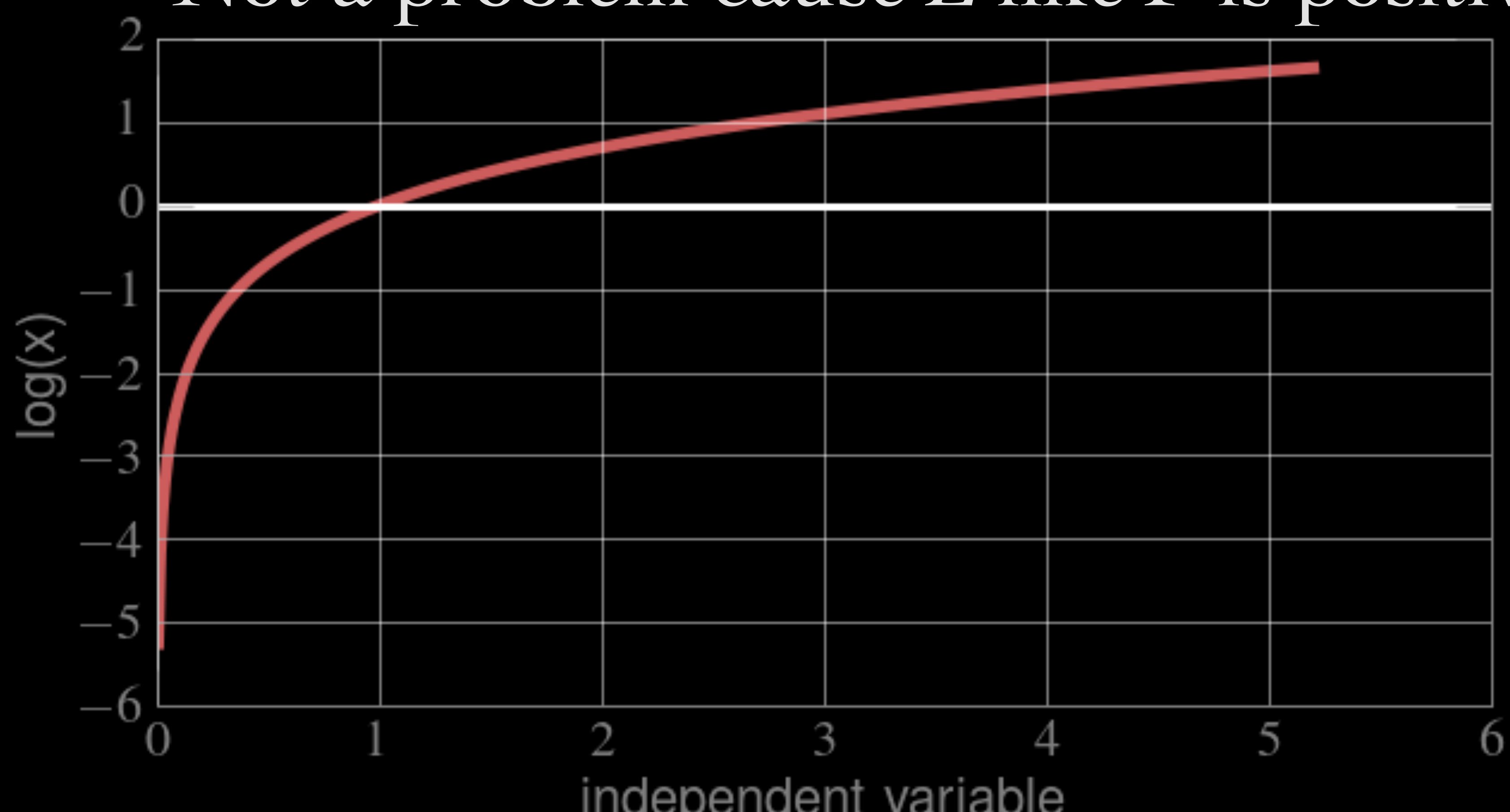
SUPPORT : (0: ∞]



Logarithm: MONOTONICALLY INCREASING

SUPPORT : (0: ∞]

Not a problem cause L like P is positive defined



Probability $$N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

log Likelihood $$\log\left(\mathscr{L}_{(\mu,\sigma)}(\vec{x})\right) \sim \log\left(\frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\Sigma_i (x_i-\mu)^2}{2\sigma^2}}\right)$$

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

log Likelihood $\log(\mathscr{L}_{(\mu,\sigma)}(\vec{x})) \sim$

$$\log\left((2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_1^n (x_i-\mu)^2\right)\right)$$

Probability

$$N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

log Likelihood

$$\ell(\mu,\sigma)(\vec{x}) \sim$$

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_1^n (x_i-\mu)^2$$

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

log Likelihood $\ell(\mu,\sigma)(\vec{x}) \sim$

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{1}^{n}(x_i-\mu)^2$$

Probability

$$N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

log Likelihood

$$\ell(\mu,\sigma)(\vec{x}) \sim$$

$$-\cancel{\frac{n}{2}\log(2\pi)} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{1}^{n}(x_i-\mu)^2$$

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

max log Likelihood

$$\ell(\mu,\sigma)(\vec{x}) \sim$$

$$\cancel{-\frac{n}{2}\log(2\pi)} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_1^n (x_i-\mu)^2$$

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

max log Likelihood $\ell_{(\mu^*,\sigma^*)}(\vec{x}) = max(\ell_{(\mu,\sigma)}(\vec{x}))$

CUSP

Probability $N(\mu,\sigma) \sim \Pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

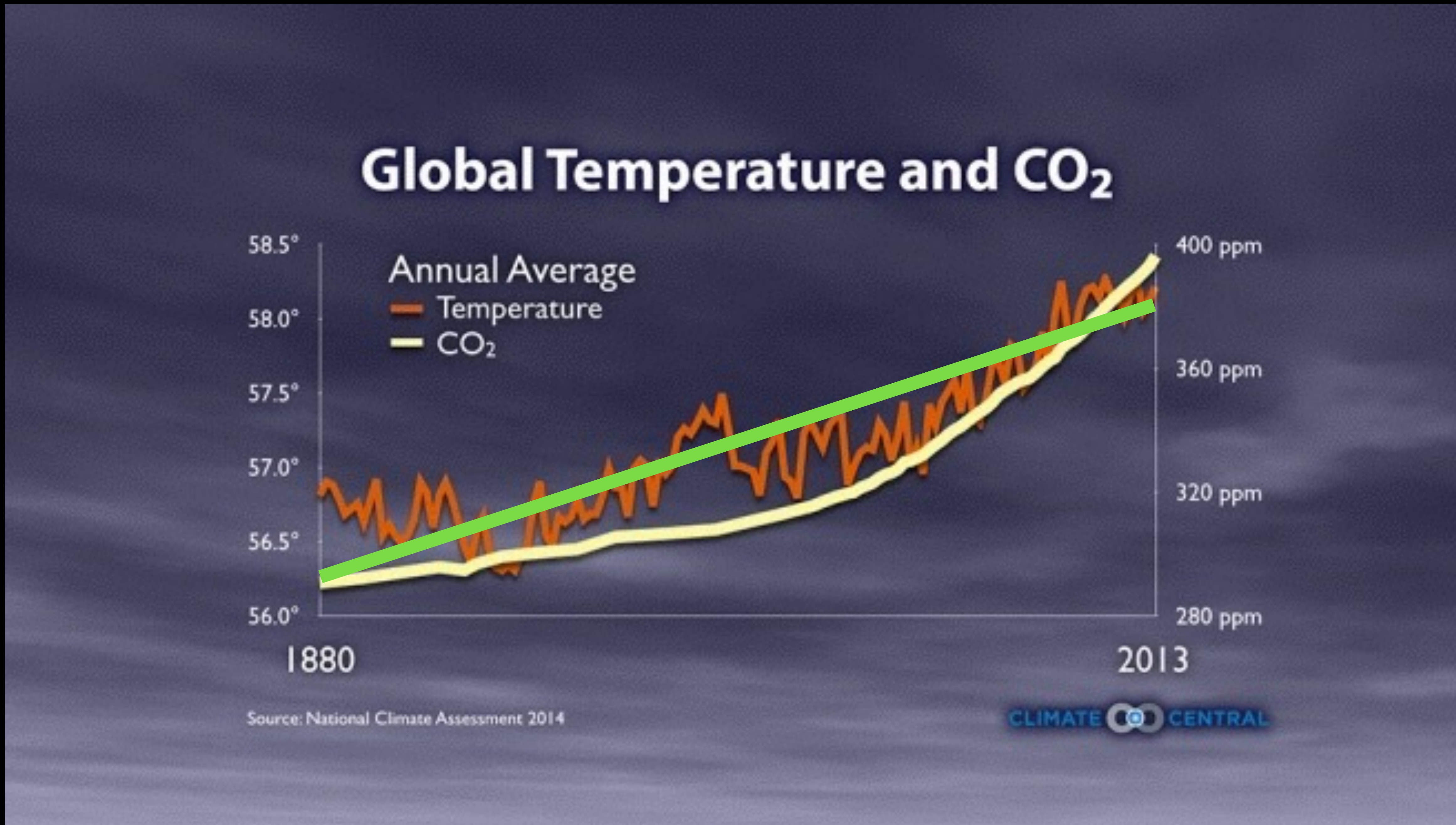
max log Likelihood $\frac{d\ell_{(\mu,\sigma)}(\vec{x})}{d(\mu,\sigma)} = 0$

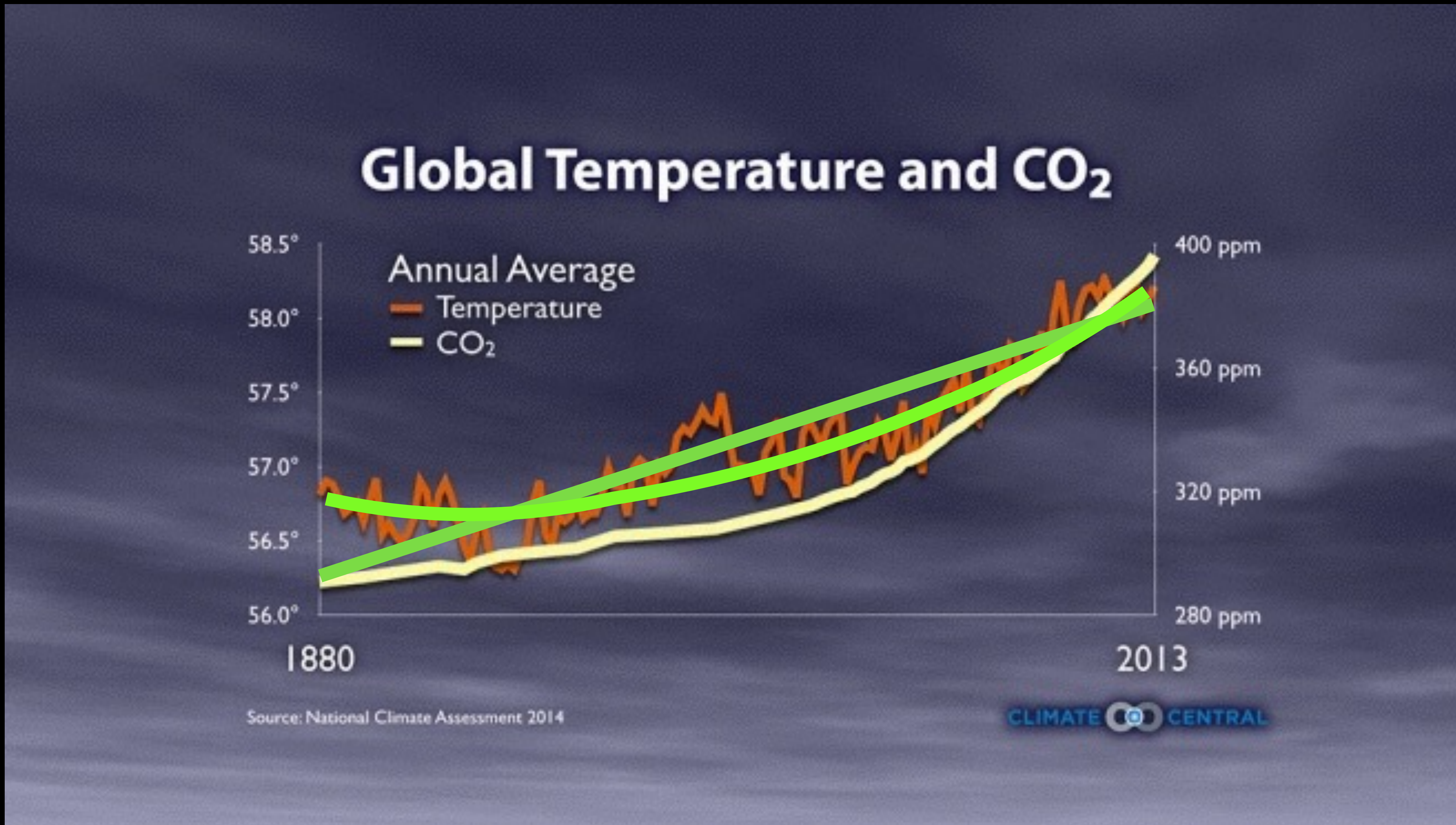
$$LR = -2 \log_e \frac{\max L(\text{model } 1)}{\max L(\text{model } 2)}$$

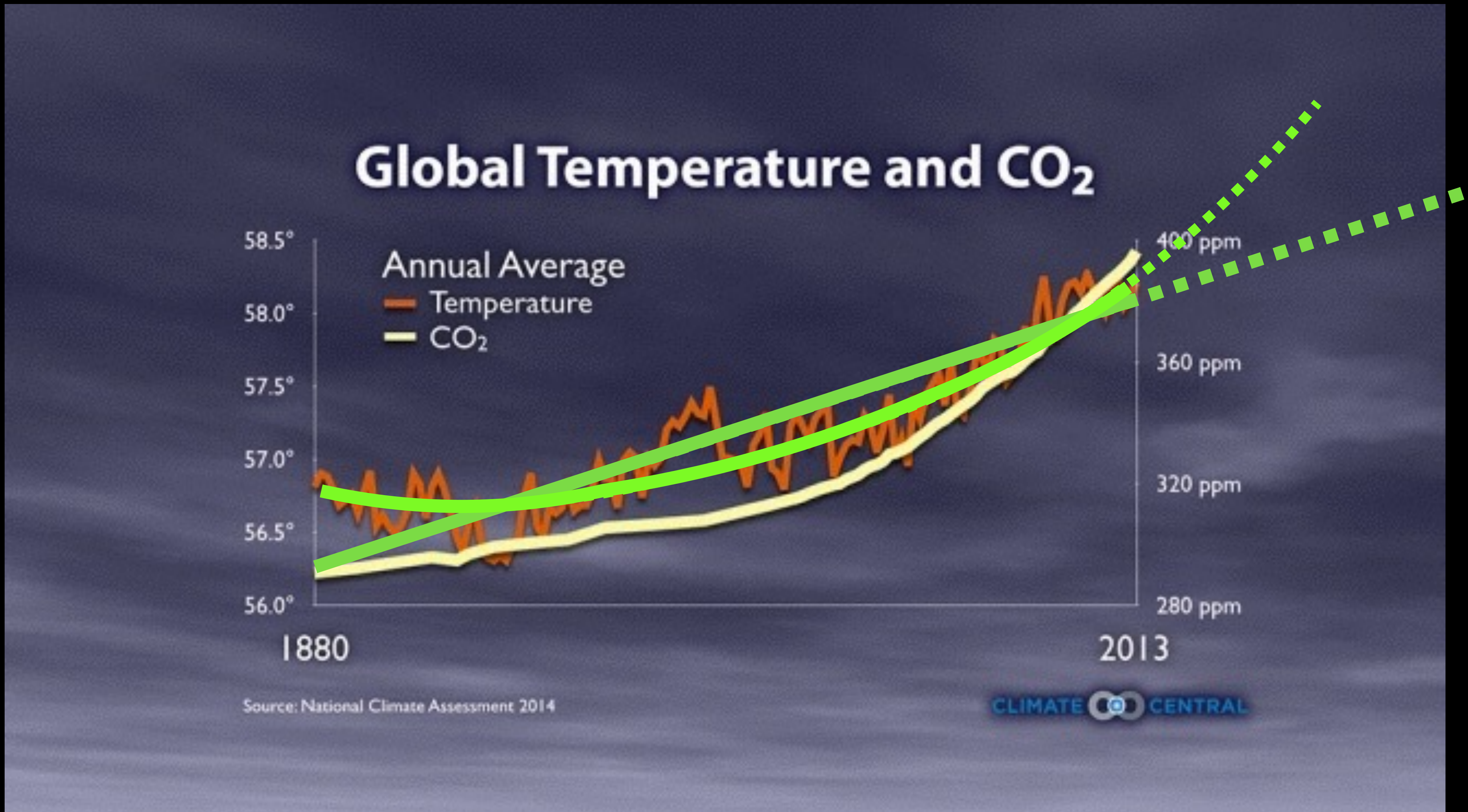
This statistic is <u>chi-squared distributed</u>

$$LR = -2 \log_e \frac{\max L(\text{model } 1)}{\max L(\text{model } 2)}$$

This statistic is chi-squared distributed with degrees of freedom equal to the difference in the number of degrees of freedom between the two models (i.e., the number of variables added to the model).


Global Temperature and CO2
Annual Average
Temperature
CO2
58.5°
58.0°
57.5°
57.0°
56.5°
56.0°
400 ppm
360 ppm
320 ppm
280 ppm
1880
2013
Source: National Climate Assessment 2014
CLIMATE CENTRAL


Global Temperature and CO_2
Annual Average
Temperature
CO_2
58.5°
58.0°
57.5°
57.0°
56.5°
56.0°
400 ppm
360 ppm
320 ppm
280 ppm
1880
2013
Source: National Climate Assessment 2014
CLIMATE CENTRAL



V: Likelihood and Regression Models

jupyter

nrg buildings notebook

Homework:

ENERGY - SIZE building modeling:
follow in class instructions

MUST KNOWS:

- How to minimize fit parameters (OLS, WLS)
- goodness of fit tests
- R^2 , χ^2 , adjusted R^2 , reduced χ^2 , likelihood, Likelihood ratio test

Resources:

Sarah Boslaugh, Dr. Paul Andrew Watters, 2008
Introduction to General Linear Regression (Chap 12 in most versions)
https://books.google.com/books/about/Statistics_in_a_Nutshell.html?id=ZnhgO65Pyl4C

David M. Lane et al.
Introduction to Statistics (XVIII)
regression : Chapter 14
http://onlinestatbook.com/Online_Statistics_Education.epub
http://onlinestatbook.com/2/index.html