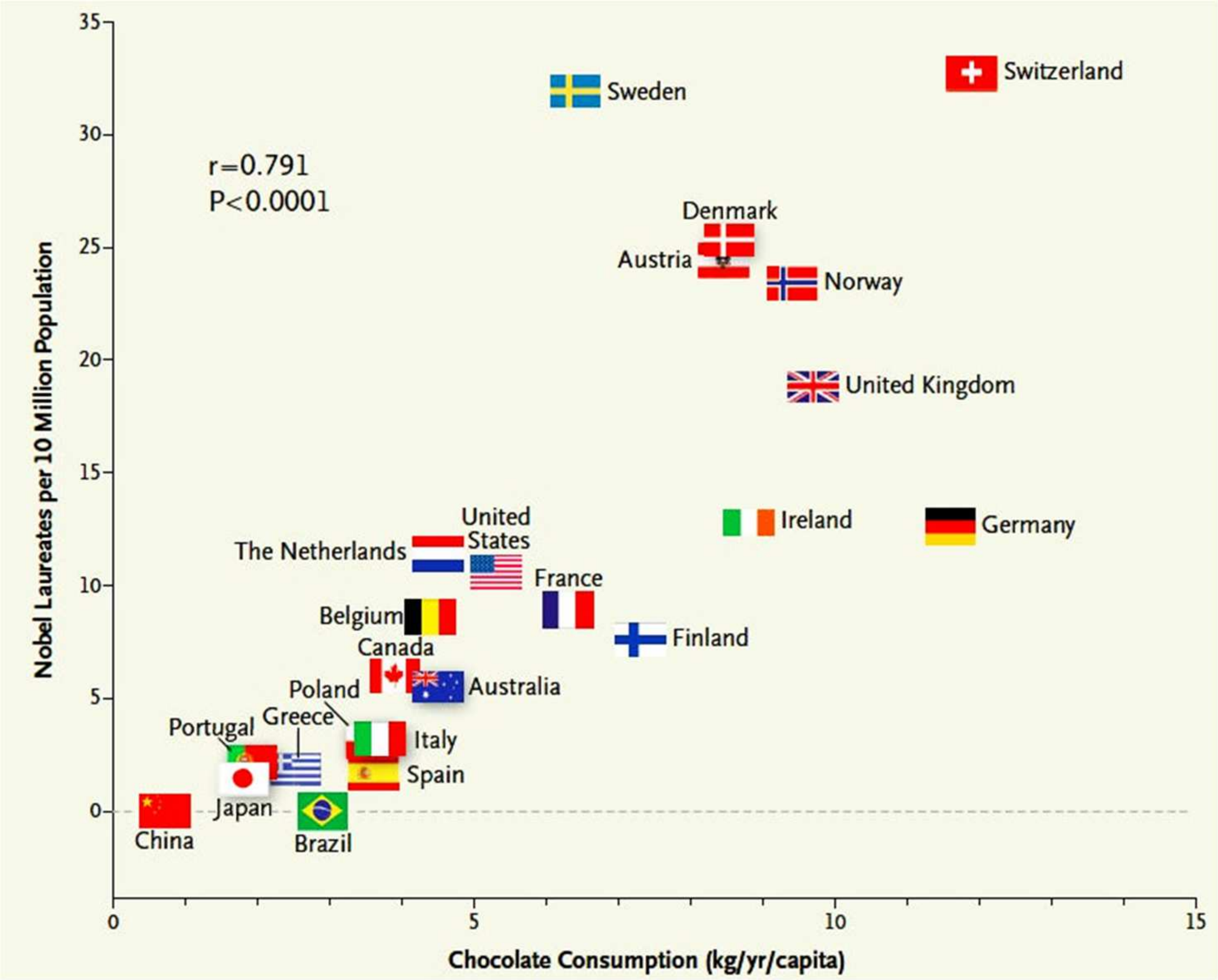
_Want more Nobel prizes?



Linear Regression

_Marco Loog

Past, Present, Future

_Previous focus largely on classification

_Today linear regression

_Tomorrow mainly classification again

With a focus on linear classifiers

What is Regression?

_Examples of prediction problems where you may not be interested in a class?

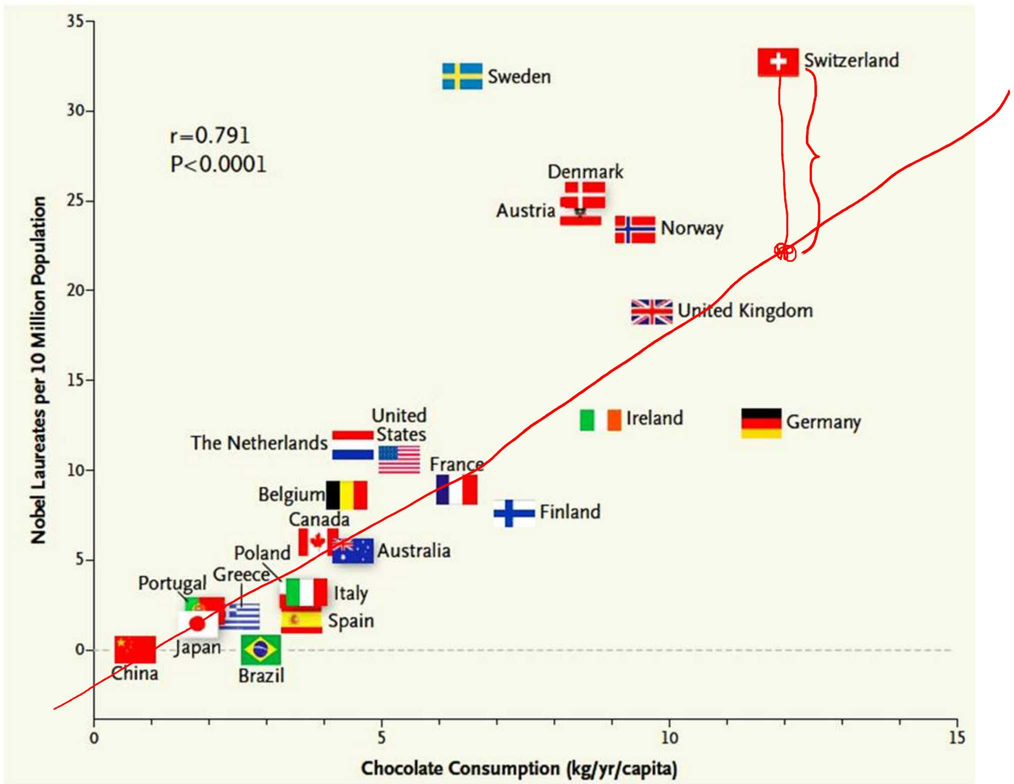
Input-Output and Error Measure

_Given $p(x, y)$

Distribution over input-output

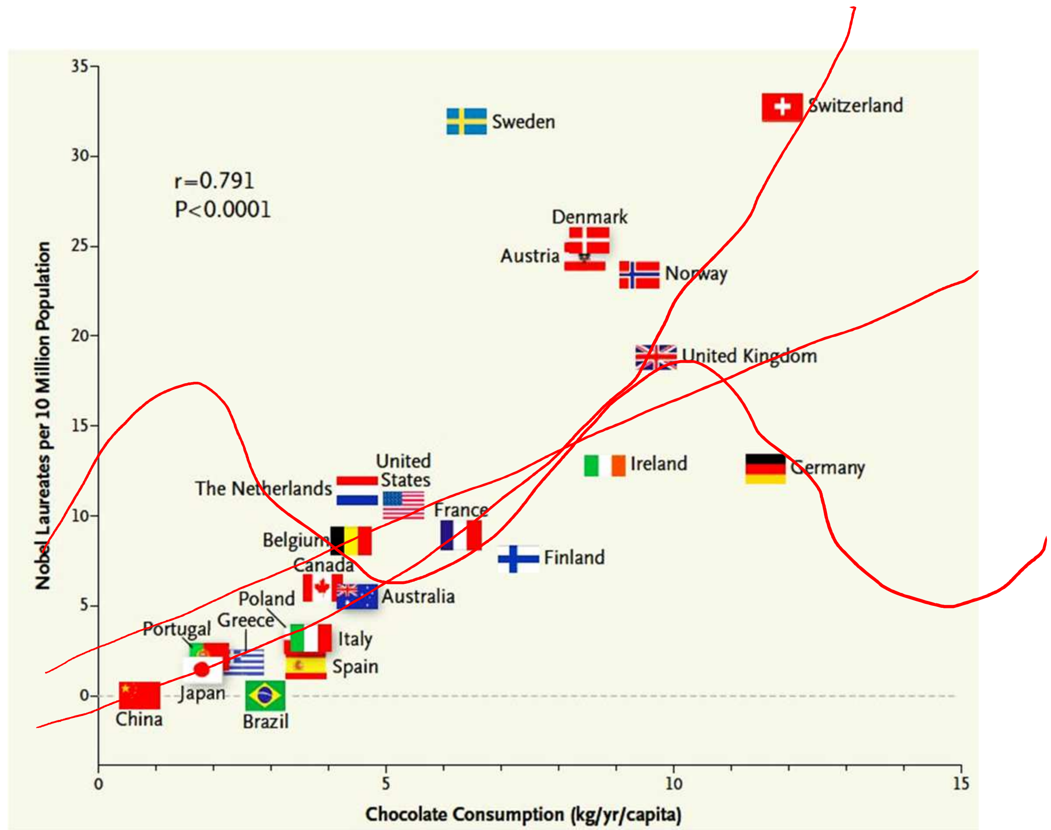
_Function $f(x)$

_How to measure goodness of fit?



Model Assumption

- Given example of (chocolate,prizes)
  - input = chocolate
  - output = prize
- What functions to consider?



Ingredients

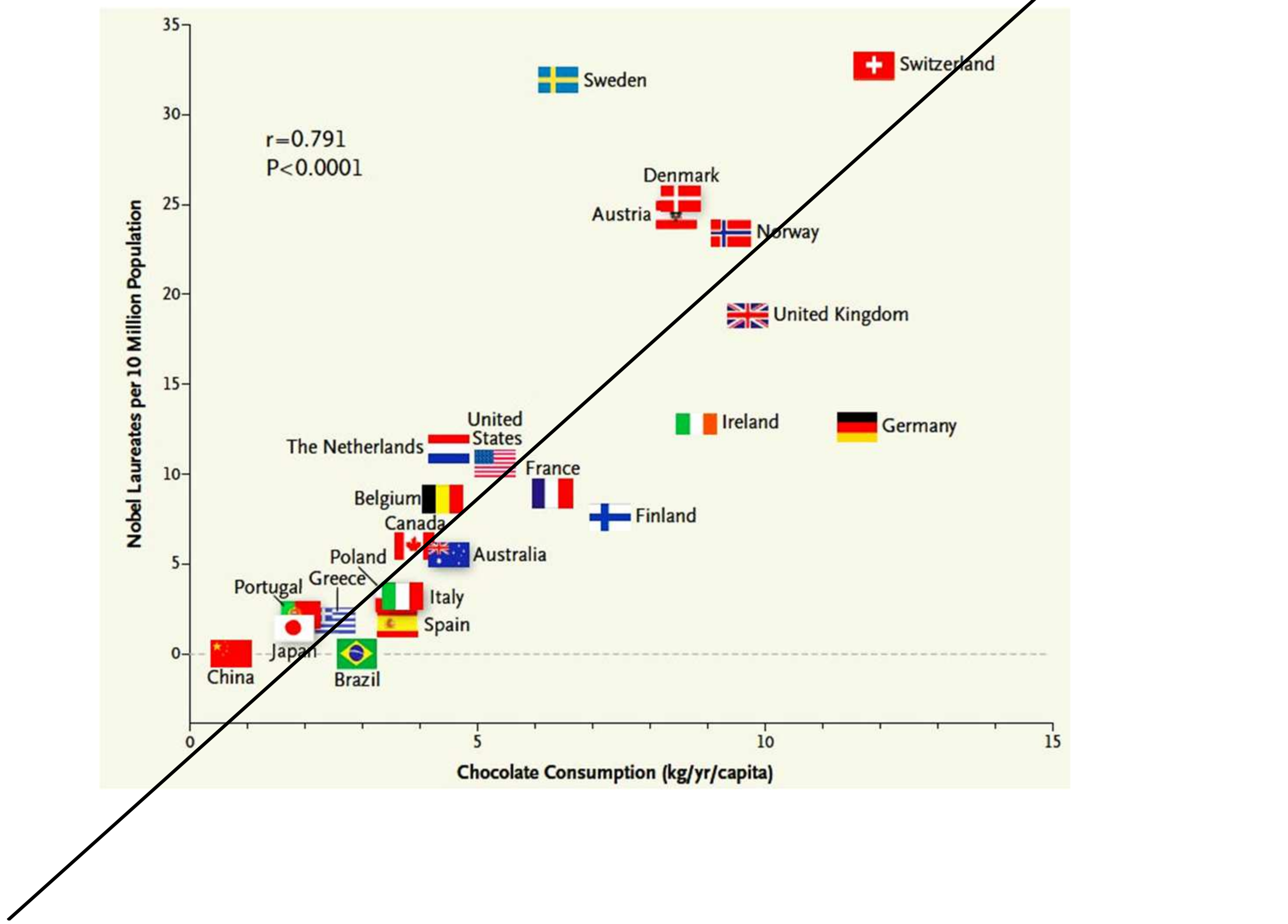
_Model

Will look at linear models

_Fitting function

Squared loss

Probabilistic



So…

_Regression aims to minimize expected squared loss

$$\int (f(x) - y)^2 p(x, y)dxdy$$

Other losses possible of course

_We do not know p

_We need to assume a model for f

Squared Loss

_Risk of interest and "Bayes regression function"?

x is fixed

at x we have the density $p(y|x)$

so optimal $f(x)$ minimizes $\int (f(x)-y)^2 \, p(y|x) \, dy$

take $\frac{d}{df(x)}$ and set to 0 $\Big\}$ so $f(x) = \mathbb{E}[y|x] = \int y \, p(y|x) \, dy$

Least Squares Linear Regression

_Assuming linearity…

Given N iid input-output pairs (x_i, y_i)

Find the w that minimizes

$f(x) = w^T x$

$$\sum_{i=1}^{N} (w^T x_i - y_i)^2 = \|Xw - Y\|^2$$

Least Squares Linear Regression

_Assuming linearity…

Given N iid input-output pairs (x_i, y_i)

Find the w that minimizes

$f(x) = w^T x$

$$\sum_{i=1}^{N} (\underline{w^T x_i} - y_i)^2 = \|Xw - Y\|^2$$

$$\sum_{i=1}^{N}(w^T x_i - y_i)^2 = \|Xw - Y\|^2$$

_Let's solve this for 1D inputs…

$$\frac{d}{dw}\sum (wx_i - y_i)^2 = \sum 2(wx_i - y_i)x_i = 0$$

$$\sum 2wx_i^2 - 2x_i y_i = 0$$

$$w\sum x_i^2 = \sum x_i y_i$$

$$w = \frac{\sum x_i y_i}{\sum x_i^2}$$

Note : Intercept / Bias

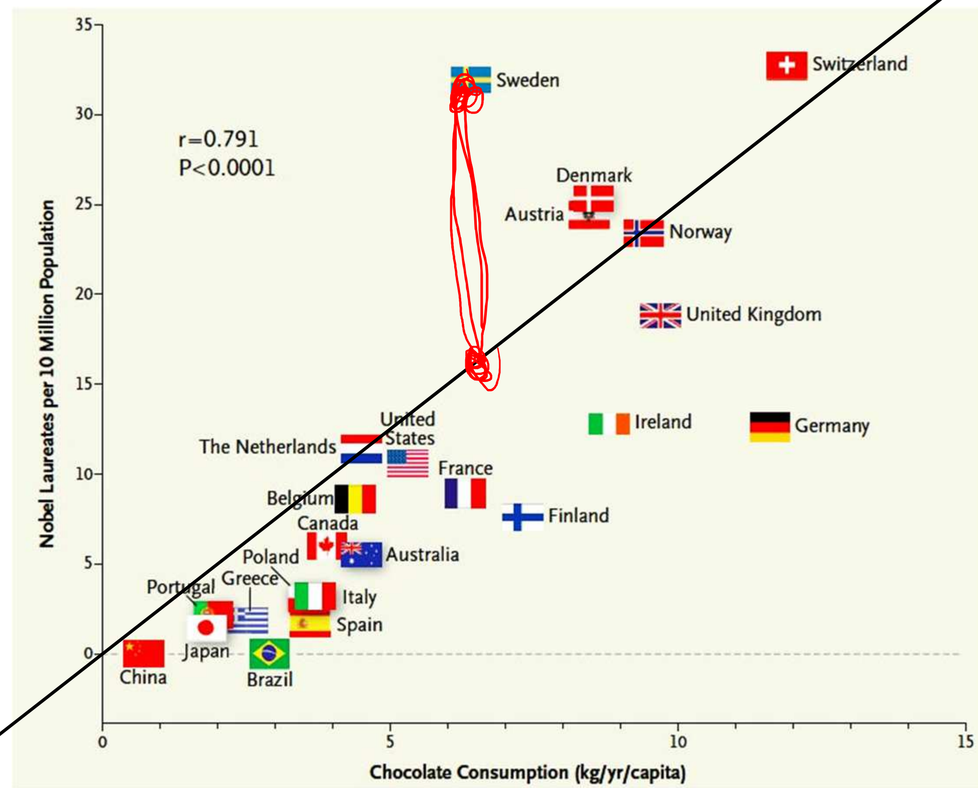
$x' = \begin{pmatrix} x \\ 1 \end{pmatrix}$

$x' = [x; 1]$

$w^T x' = w^T x + w_d$

_$w^T x$ always goes through 0 for input 0

How do we fix this?



Q? / Recap / Remainder

_Regression is for ordered / continuous outputs

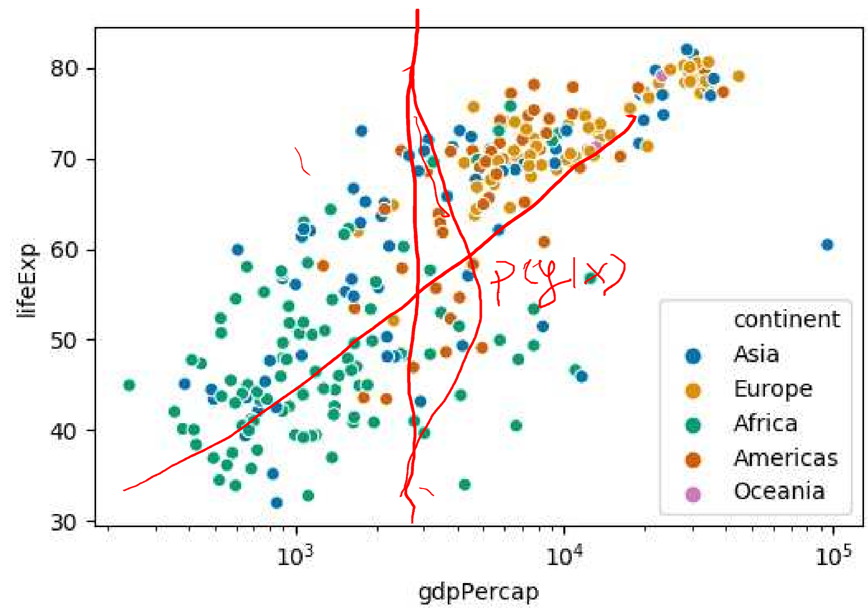
$$\sum_{i=1}^{N} (w^T x_i + w_0 - y_i)^2$$

Probabilistic extension
Simple prior knowledge
“Nonlinear” model

Extension to Probabilistic Model

_But why?

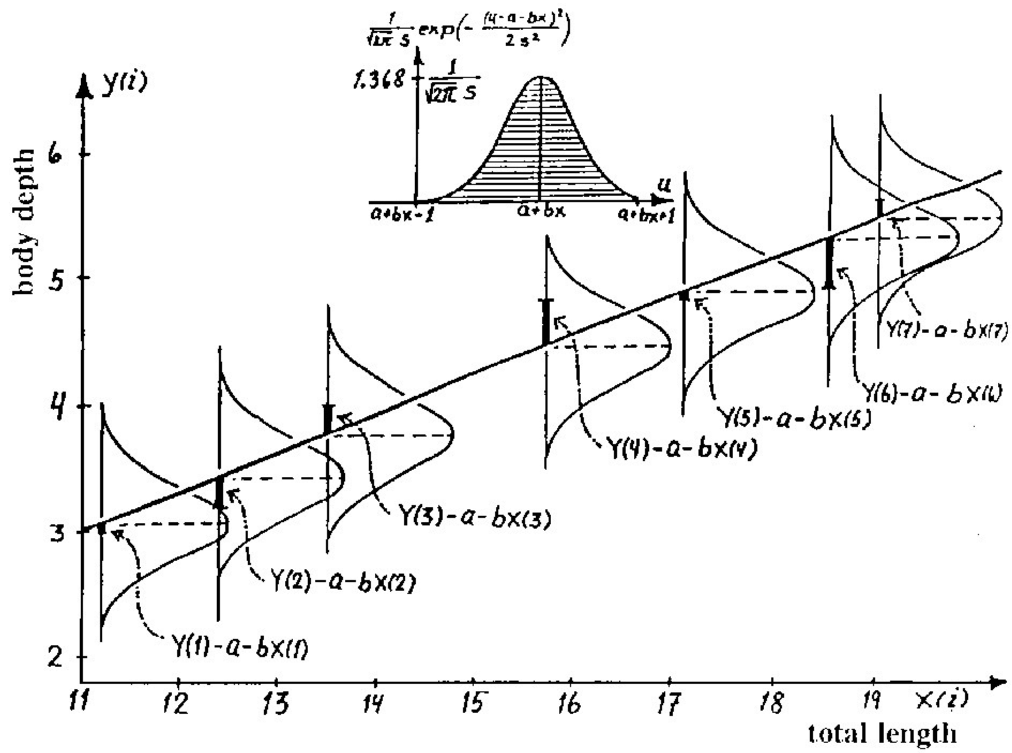
Model spread in prediction
Express confidence



Extension to Probabilistic Model

_How to?

One possibility is to assume conditional for Gaussian $p(y|x)$



How To

_Conditional at x : $p(y|x) = N(y|w^T x, \sigma^2)$
_Fit to data by maximizing (conditional) likelihood

$$\prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2)$$

What are the parameters to optimize?
Depends on what the model assumes…

$$\prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2)$$

_Let's fit it

Assume σ known

$$\sum \log N(y_i|w) = \sum_i \left[C + -\frac{1}{2\sigma^2}(w^T x_i - y_i)^2 \right]$$

$$\hat{w}_{ML} = (X^T X)^{-1} X^T Y$$

$$\prod_{i=1}^{N} N(y_i | w^T x_i, \sigma^2)$$

_Let's fit it

Assume w known

$$\boxed{\frac{1}{M}\sum_{i=1}^{M}(\hat{\mu}-a_i)^2}$$

$$\hat{\sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\underbrace{(\hat{w}_{ML}^T x_i - y_i)}_{\hat{w}_{ML}^T x_i - y_i}{}^2$$

Q? / Recap / Next Topic

_Can reinterpret standard linear regression in terms of a probabilistic model

_Now : important way of incorporating prior knowledge [more on this in Week 5]

Rough Idea

- Estimate average height in football team
- What do you do in case of 0 observations?

Maximum a Posteriori Estimation

_One way of combining a prior information with actual data : take likelihood × prior

$$p(\text{data}|\theta)p(\theta)$$

_MAP estimate obtained by maximizing for θ

So, think about how you would approach team height estimation…

Generic Prior in Regression

_Assume that w is [relatively] close to 0
_More specifically take prior $N(w|0, \alpha I)$ [α = fixed!]
_MAP estimate $\hat{w}_{\mathrm{MAP}}$ maximizes

$$\left(\prod_{i=1}^{N} N(y_i | w^T x_i, \sigma^2)\right) N(w|0, \alpha I)$$

You should be able to solve this [at least for 1D case, σ fixed]

Generic Prior in Regression

_MAP estimate $\hat{w}_{\text{MAP}}$ maximizes

$$\left(\prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2)\right) N(w|0, \alpha I)$$

_Solution for this specific choice [with σ fixed]

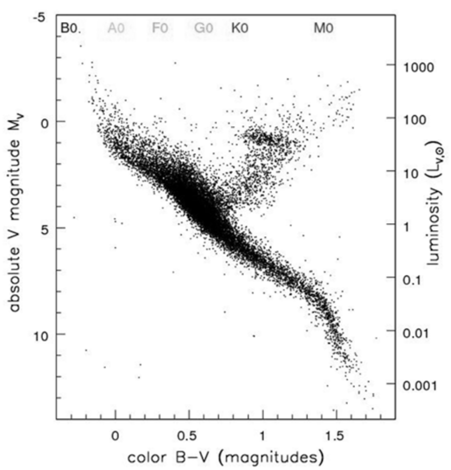
$$\hat{w}_{\text{MAP}} = \left(X^T X + \frac{\sigma^2}{\alpha} I\right)^{-1} X^T Y$$

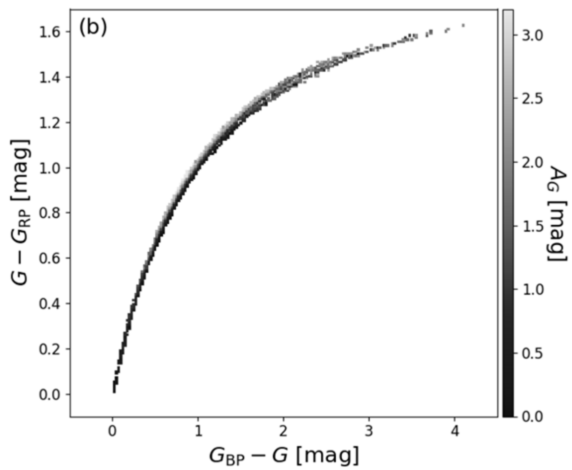
Next : Nonlinear Relations…

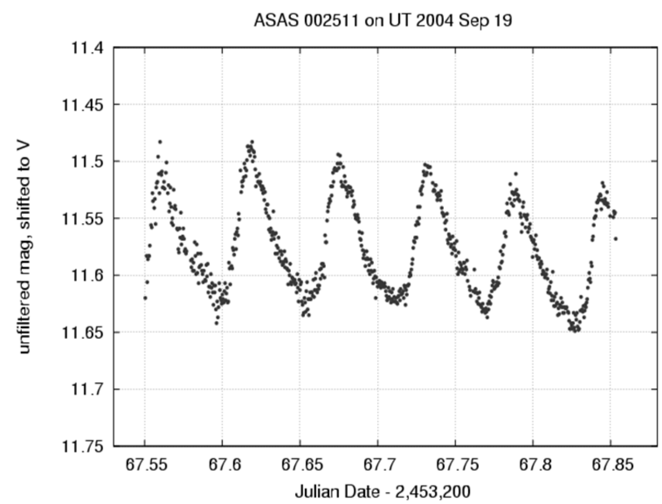
_Often variables relate in a nonlinear way

_$E = mc^2$, $G = \frac{m_1 m_2}{r^2}$, etc.

_What can we do?







Feature Transformations

_Nothing prevents inventing own combinations

_Already added constant for intercept / bias / offset

_Why stop there?

With $x \in \mathbb{R}^3$ a feature vector, we could add…

x_1^2, $\sin x_3$, $x_1 x_2$, etc.

[Note potential confusion with indexed samples]

$$\begin{pmatrix} x_1 \\ x_1 x_2 \\ \sin x_3 \\ x_3^x \end{pmatrix}$$

_Generally, invent mapping $\phi\colon \mathbb{R}^d \longrightarrow \mathbb{R}^D$ from d-dimensional space to new D-dimensional one

Feature Transformations

_With your choice of ϕ, new objective becomes

$$\sum_{i=1}^{N} (w^T \phi(x_i) - y_i)^2$$

Typically, model is still called linear

_Special case : polynomial regression of some order

_Relation to the kernel trick [Week 4]

Wrap-up

_Discussed regression, linear in particular
_Both squared loss formulation and probabilistic
_Extensions using prior and feature transformations

_Tomorrow we look at linear classifiers
_Think about the following :

Which linear ones did you see already?
How to use linear regression to build a linear classifier?