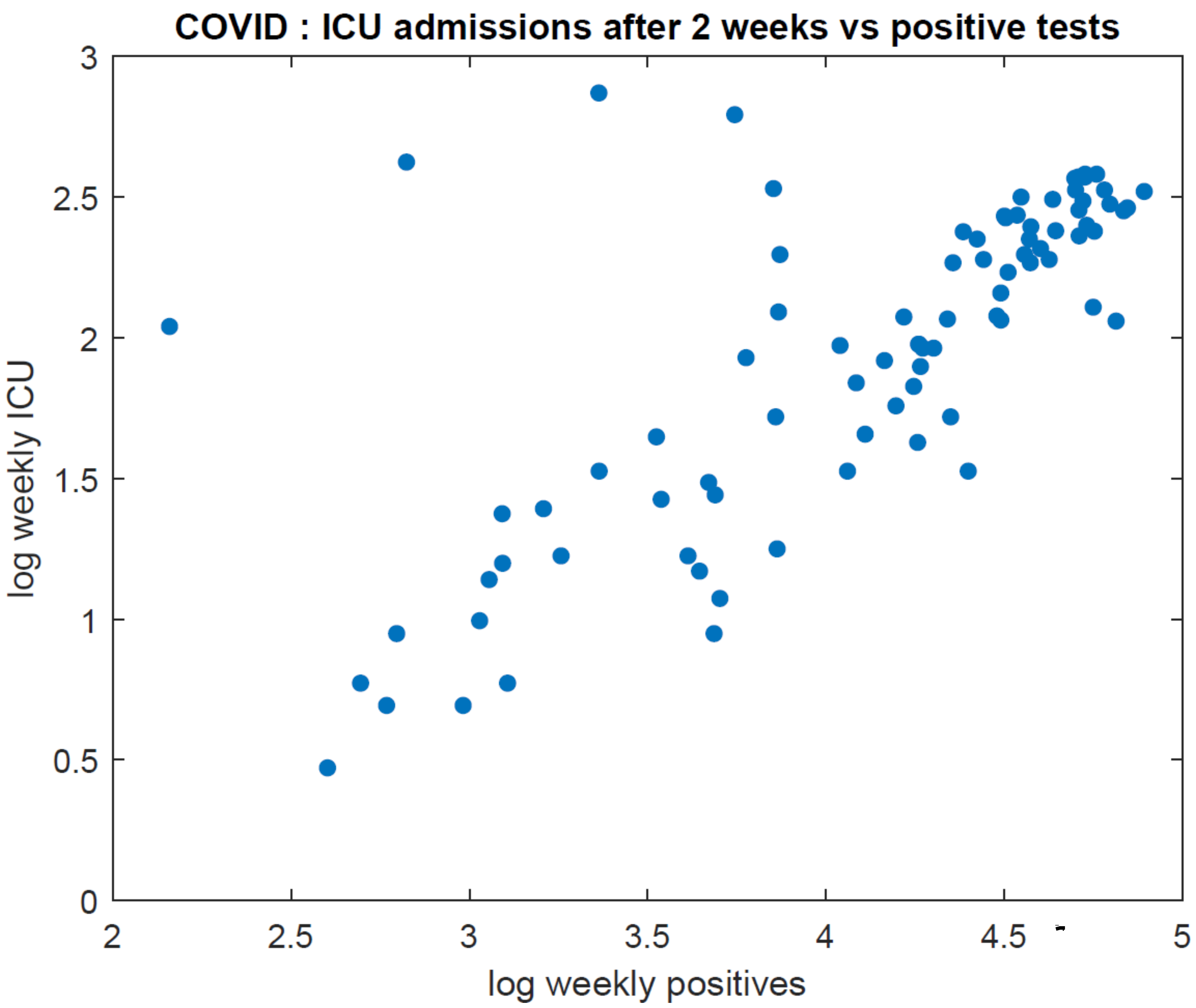
- Predict ICU admissions two weeks ahead based on positives



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

Why Regression?

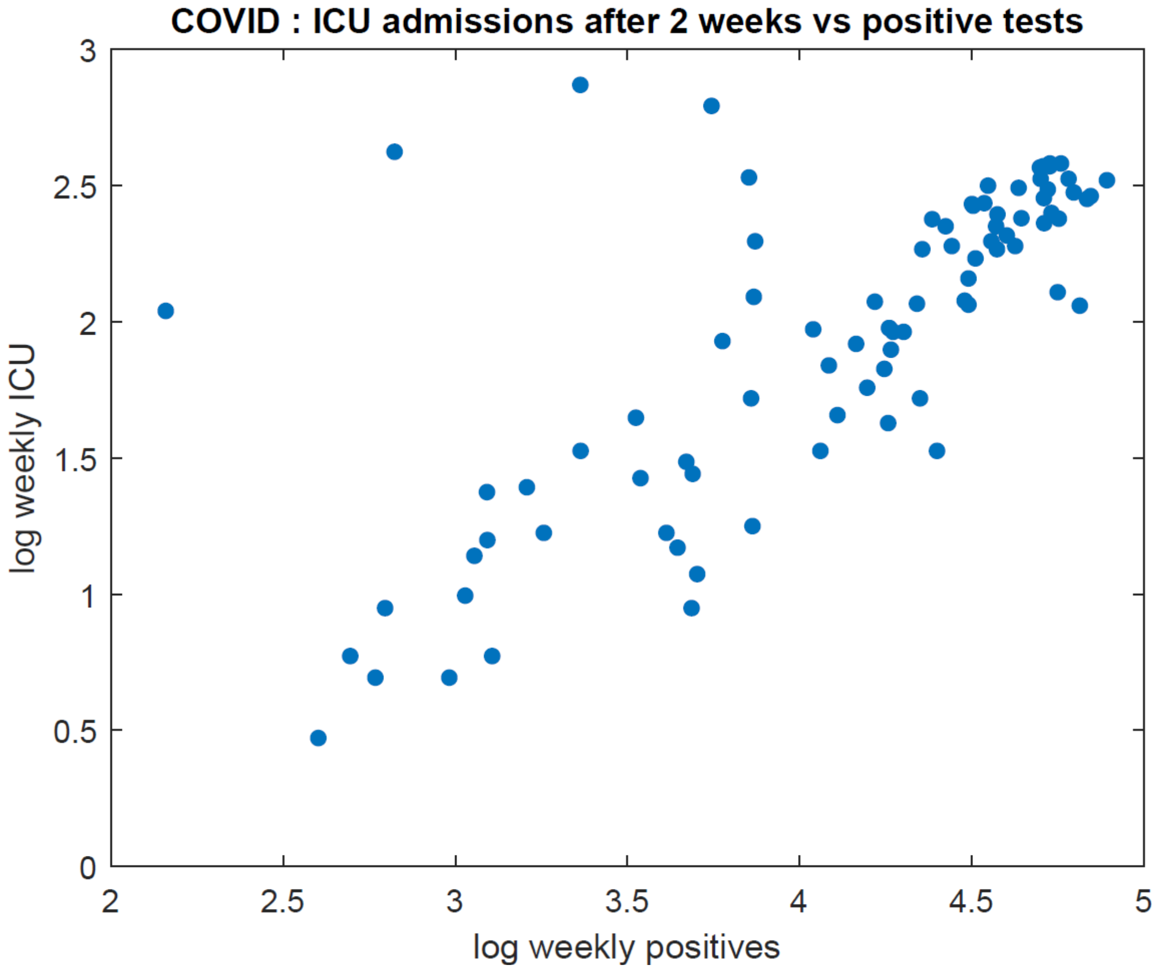
- Other examples of prediction problems where you may not be interested in a class label?

Input-Output and Error Measure

- Given input-output data

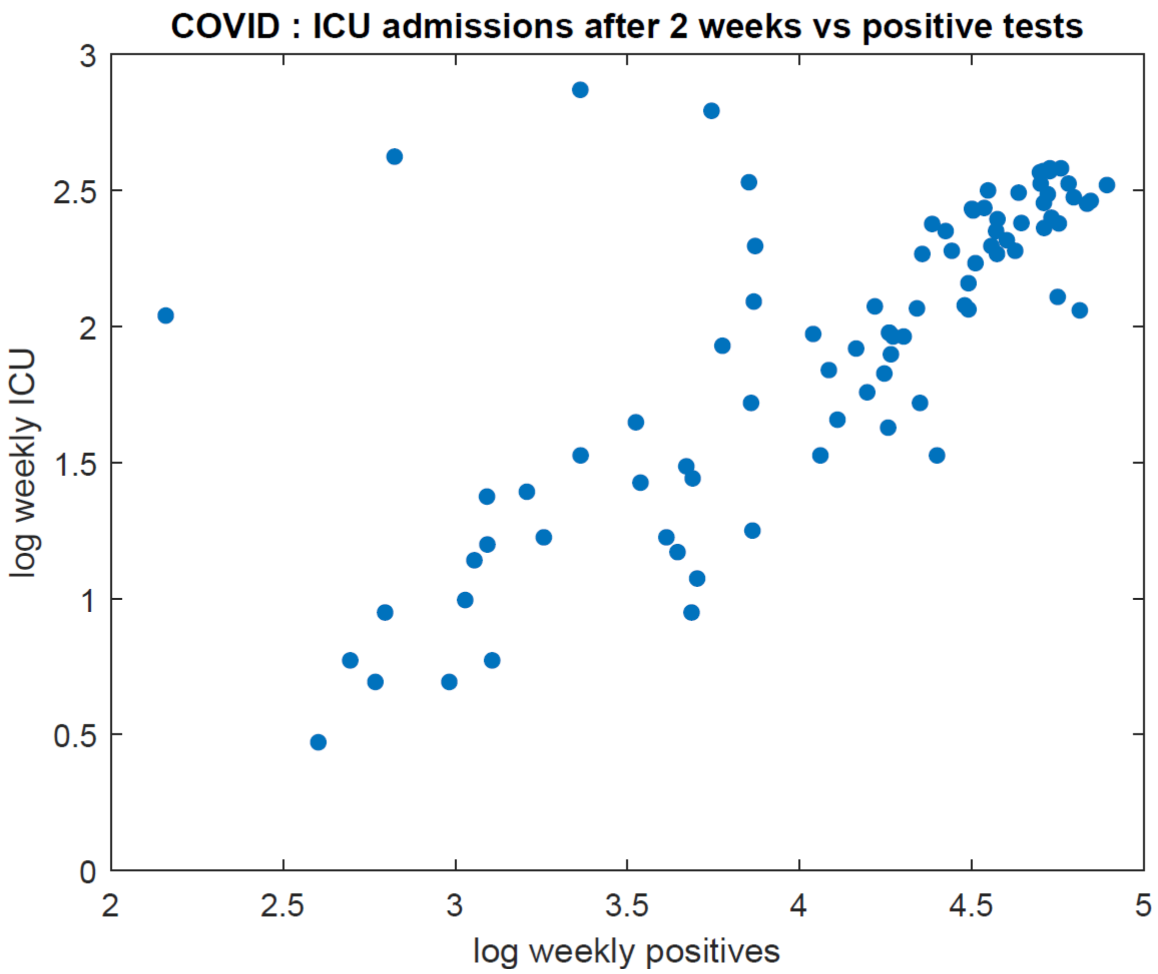
- Function $f(x)$

- How to measure goodness of fit?



Input-Output and Error Measure

- Given $p(x, y)$
  - Distribution over input-output
- Function $f(x)$
- How to calculate goodness of fit?



Taking Squared Loss…

- Risk of interest and "Bayes regression function"?

    I.e., what is the optimal solution given $p(x, y)$?

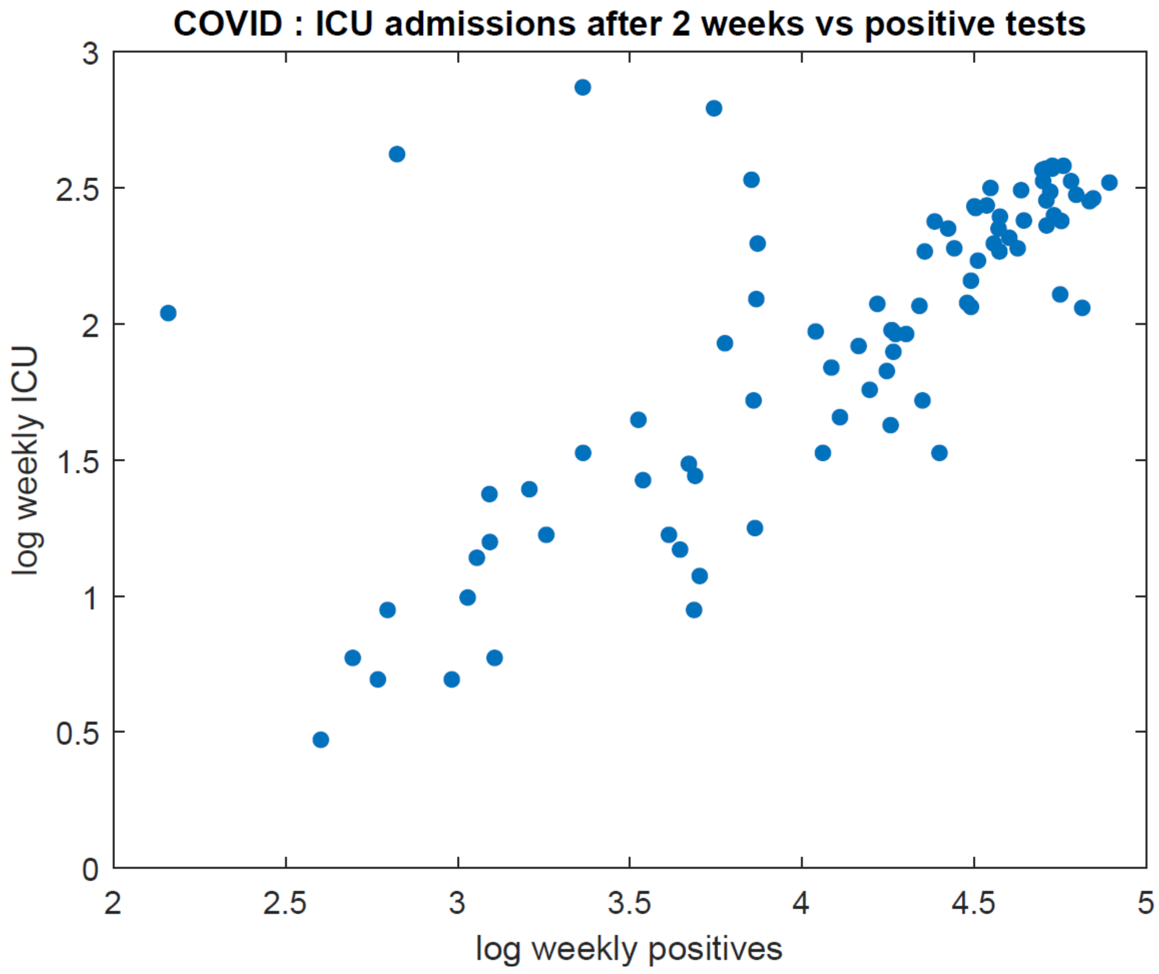
    Consider a fixed x for this…

Model Assumption

- Given examples of (positives,admissions)

  input = positives
  output = admissions

- What functions to consider?



Ingredients

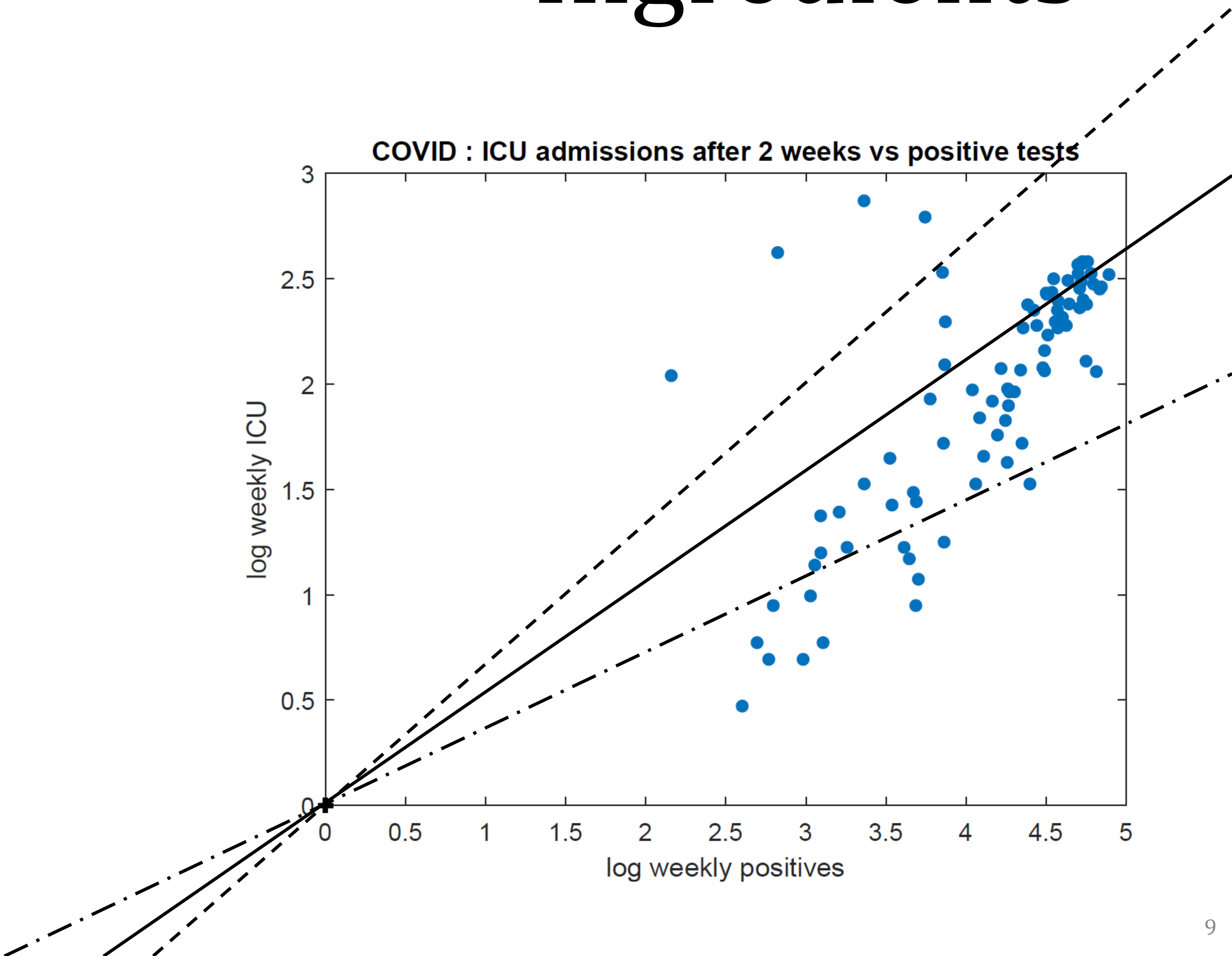
› Model

Will look at linear models

› Fitting function

Squared loss

Probabilistic



So…

- Regression aims to minimize expected squared loss

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

Other losses possible of course

- We do not know p
- We need to assume a model for f

Least Squares Linear Regression

› Assuming linearity…

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

Find the w that minimizes

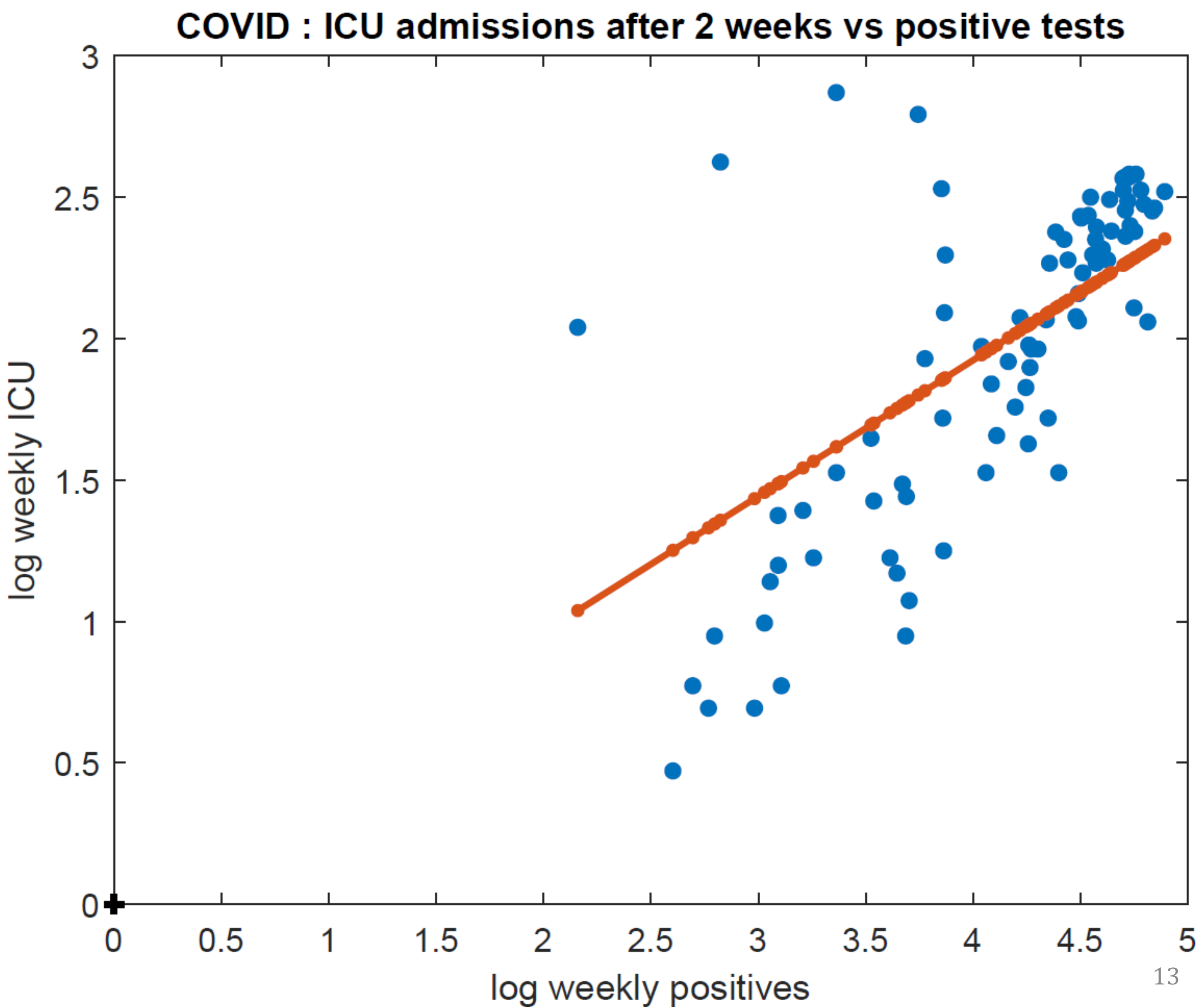
Note : input typically is multidimensional!

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

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

› Let’s solve this for 1D inputs…

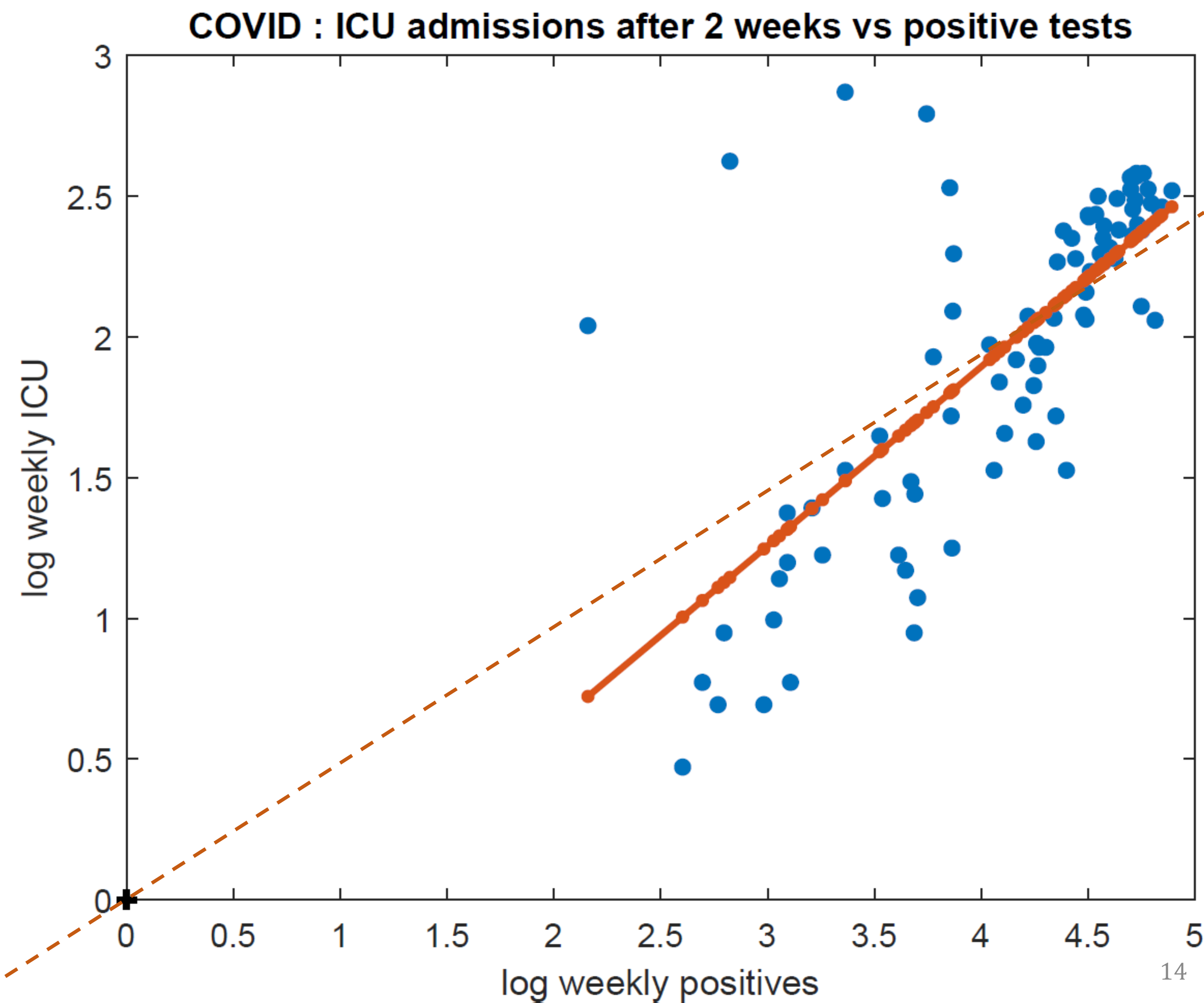
On Our Running Example



Note : Intercept / Bias

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

  How do we fix this?



Q? / Recap / Remainder

› Regression is for ordered / continuous outputs

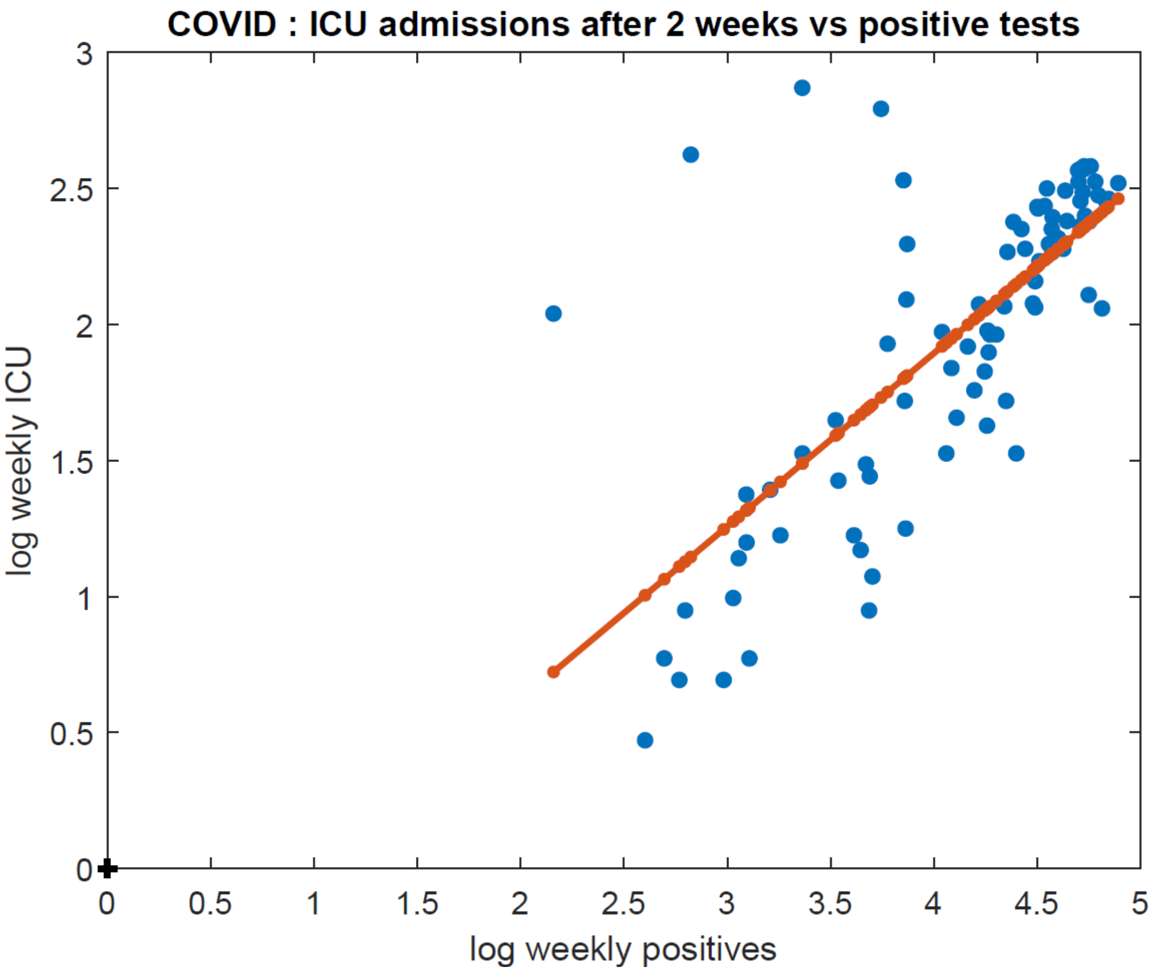
$$\sum_{i=1}^{N}(w^T x_i + w_0 - y_i)^2$$

Probabilistic extension
Simple prior knowledge
“Nonlinear” model

Extension to Probabilistic Model

› But why?

Model spread in prediction
Express confidence
Combine with other probabilistic models

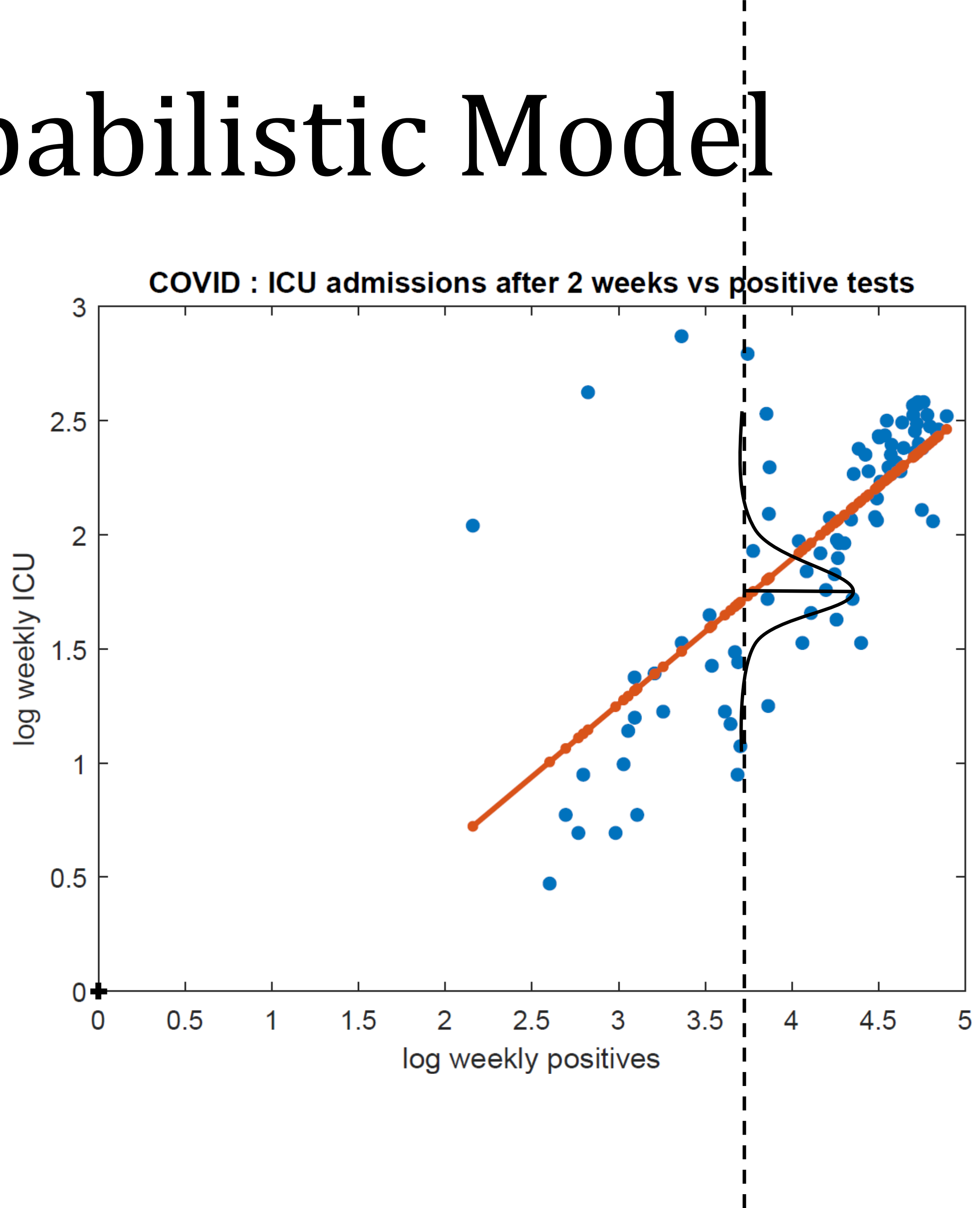


Extension to Probabilistic Model

› How to?

Again : assume a model

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



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 assuming σ known…

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

› What if we assume w known?

Q? / Recap / Further Topics

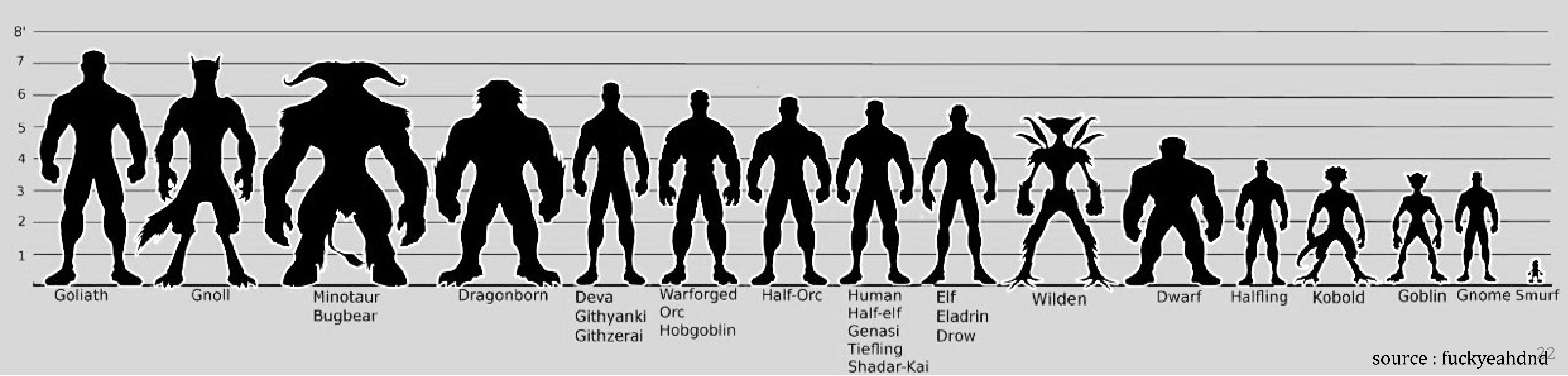
› Can reinterpret standard linear regression in terms of a probabilistic model

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

› Second : nonlinear relations [relates to Week 4]

Initial Idea

- Estimate average student height in specific ML class
- What do you do in case of 0 observations?



source : fuckyeahdnd

Maximum a Posteriori Estimation

› One way of combining a prior information with actual data : take likelihood × [so-called] prior

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

› MAP estimate obtained by maximizing for θ

So, think about how you would approach ML student height estimation…

Generic Prior in Regression

- Assume that w is [relatively] close to 0
- More specifically take prior $N(w|0, \alpha I)$ [$\alpha =$ fixed!]
- MAP estimate $\widehat{w}_{\mathrm{MAP}}$ 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 $\widehat{w}_{\mathrm{MAP}}$ 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]

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

Behavior?

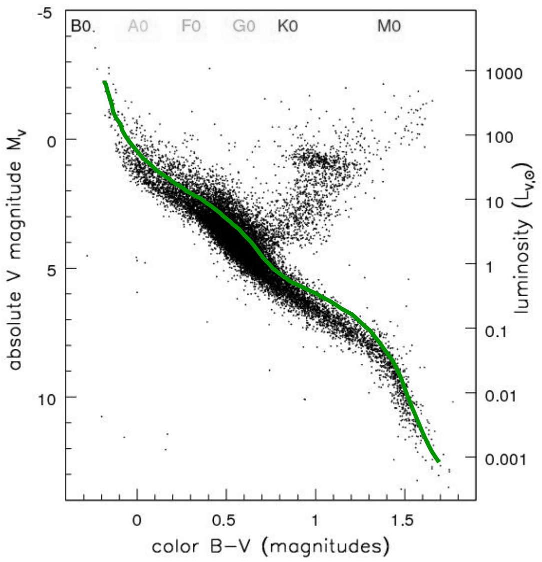
› Solution for this specific choice [with σ fixed]

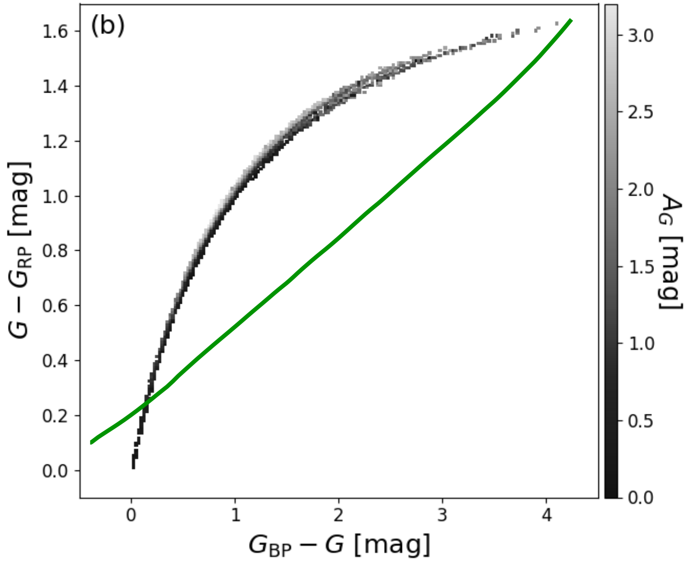
$$\hat{w}_{\mathrm{MAP}} = \left(X^T X + \frac{\sigma^2}{\alpha} I\right)^{-1} X^T Y$$

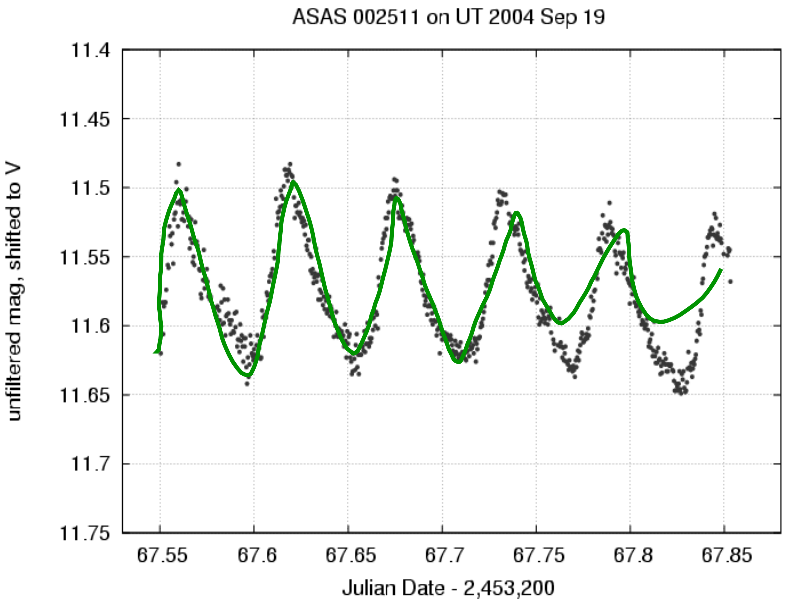
What if $\alpha \to \infty$?
What if $\alpha \downarrow 0$?
Makes sense?

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

Feature Transformations

› With your choice of ϕ, new objective becomes

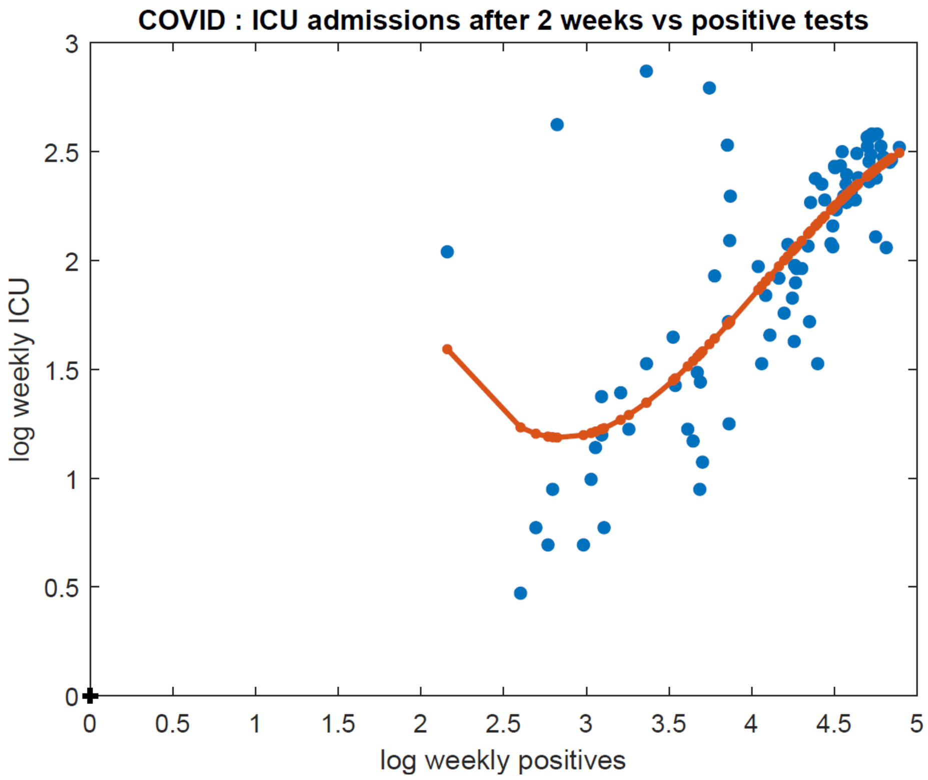
$$\sum_{i=1}^{N} (w^T \phi(x_i) - y_i)^2$$

Typically, model is still called linear

› Special case : polynomial regression of some order

› Relation to the kernel trick [Week 4]

Feature Transformations



- 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?