W1D3 Outro: Model Fitting

Kunlin Wei, PhD

neuromatch academy

Kunlin Wei

- Professor at the School of Psychological and Cognitive Sciences, Peking University, Beijing, China
- Interests: human motor control and learning, motor behaviors in general
- Tools: psychophysics, virtual reality, modeling, brain stimulations.
- Lab website: http://www.psy.pku.edu.cn/weikunlin/english/home/index.htm

What we learned today (W1D3)

- Model fitting by MSE/MLE
- Linear regression as a starting point
- Parameter uncertainty estimated by bootstrapping
- Bias-variance tradeoff
- Model selection by cross-validation

What in this outro

1. **Review:** MLE and its limitations, side-by-side comparisons of cross-validation and Bootstrapping
2. **Example**: how model fitting is used for research
3. **What's next:** go beyond simple linear models

(1) Quick review and more

Limitations of the Frequentist Framework

Crossvalidation VS. Bootstrapping

MLE is a frequentist framework

- Deal with generative models, i.e., any model/process that is assumed to give rise to some data.
- By maximizing the likelihood, $P(data|parameter)$.
- Lead to point estimates of parameters with their confidence interval.

- Interpretation in the frequentist framework: the estimated parameter is most consistent with ***the current dataset.*** And, we cannot say that the 95% CI covers the true parameter with a probability of 95%. Instead, we say that if generate samples for a large number of times, CI is the theoretical range that it will cover the true parameter for 95% of times.

Limitations of MLE

The competing, Bayesian framework:

$$\underbrace{P(parameter|data)}_{\text{Posterior}} \propto \underbrace{\boxed{P(data|parameter)}}_{\text{Likelihood}} \underbrace{P(parameter)}_{\text{Prior}}$$

Bayesian:

- Use Likelihood + Prior to estimate parameters, i.e., use sample data to update prior belief of the parameter.
- CI is related to the probability (or uncertainty) of the posterior.

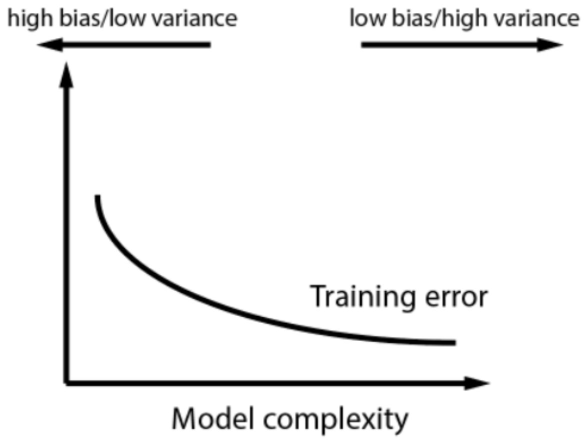
Limitations of the Frequentist approach

- Completely rely on the data, thus subject to biases of a particular random sample.
- Obtain the most likely parameters without considering prior knowledge, sometimes computationally expensive.

Cross-validation

- **The bias-variance tradeoff** in model building.
- Increase model complexity, the bias will decrease and the variance will increase. Decrease the model complexity, the bias will increase and the variance will decrease.

- Solution: use **cross-validation** to test the model. e.g., 10-fold CV, divide the dataset.
- Test error shows model performance.



Bootstrapping

Recap:

1) obtain an independent data set by resampling the original data set with replacement;
2) compute model parameters;
3) do it many times;
4) produce the confidence interval for the model parameter.

A linear regression example

	Contrast	Spike
1	576	3.39
2	635	3.3
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3
8	661	3.43

θ_1 = 0.053

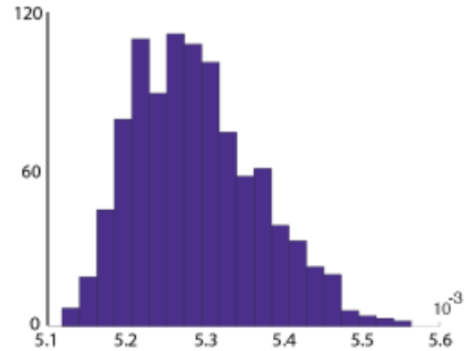
#1

1	576	3.39
8	661	3.43
6	580	3.07
3	558	2.81
4	578	3.03
8	661	3.43
5	666	3.44
6	580	3.07

#2

3	558	2.81
6	580	3.07
2	635	3.3
2	635	3.3
1	576	3.39
3	558	2.81
5	666	3.44
6	580	3.07

⋮



Bootstrapped θ_1

Crossvalidation VS. Bootstrapping

	Crossvalidation	Bootstrapping
Common	Both are resampling methods, computationally expensive (CPU hungry)	
Purpose	Good for estimating the model prediction errors	Good for estimating the confidence interval of model parameters.
Approach	Split the data into multiple sets, thus no overlapping between datasets.	Clone the data to create more sets, thus overlapping datasets.
Sample size	Needs a large sample size	Fine with small samples

(2) A complete example

Two ways of using a model:

1) Parameter estimation, fitting the model

2) Simulation, making predictions from the model

Steps to build and fit a model

1. Conceptualize a model, based on prior knowledge/hypothesis
2. Mathematically formulate it
3. Model fitting, **estimating model parameters**
4. Compare **model simulations** to empirical data
5. Interpret model parameters
6. Generate new experimental predictions

An example study:
human behavior + modeling

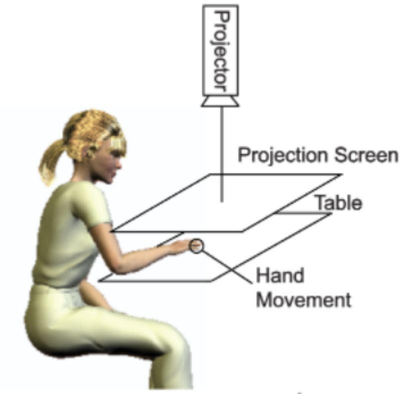
Relevance of Error: What Drives Motor Adaptation?

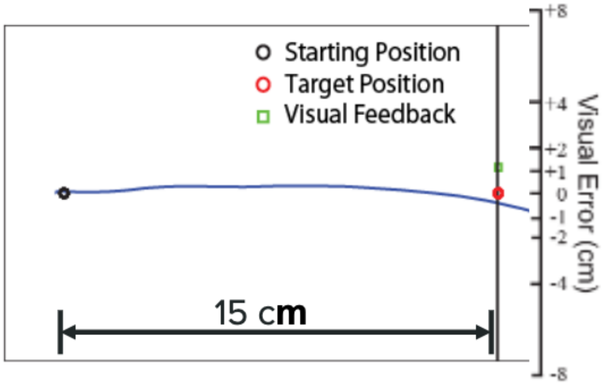
Kunlin Wei and Konrad Körding
Departments of Physiology, Physical Medicine and Rehabilitation, and Applied Mathematics, and Rehabilitation Institute of Chicago, Northwestern University, Chicago, Illinois

Submitted 8 May 2008; accepted in final form 12 November 2008

Conceptualize a research idea

A motor problem: shall I adapt to a visual error?

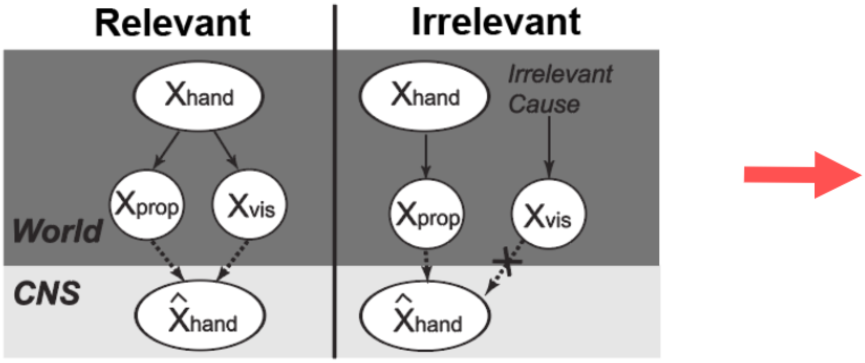


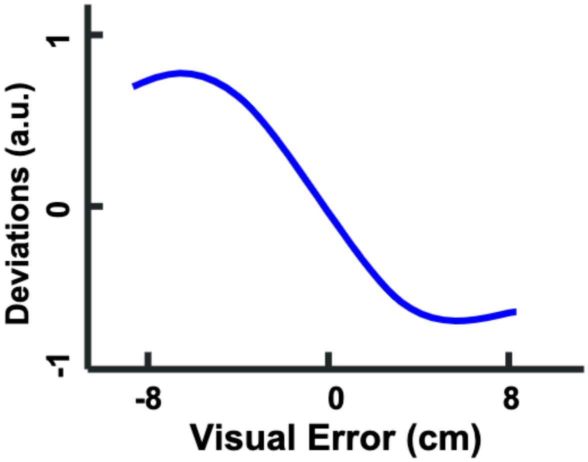


1) Small visual error → more likely self-produced error → then I shall correct it.
2) Large error → more likely caused by irrelevant factors → then I shall neglect it.

The idea shown as a graph

The motor system is doing probabilistic causal inference!





X_{vis} and X_{prop} are visual and proprioceptive cues, respectively.

Formulate the model

The final model takes ***a nonlinear form***:

$$\boxed{\hat{x}_{hand}} = \boxed{x_{vis}} \cdot \boxed{s} \cdot \frac{N(\boxed{x_{vis}}, \boxed{\sigma})}{N(x_{vis}, \sigma) + \boxed{c}}$$

The model shows how much the hand should move as a function of visual perturbation.

Observed data:

- $\hat{x}_{hand}$ is the estimated hand location at a trial
- x_{vis} is the visual perturbation applied (0,±1, ±2cm...)

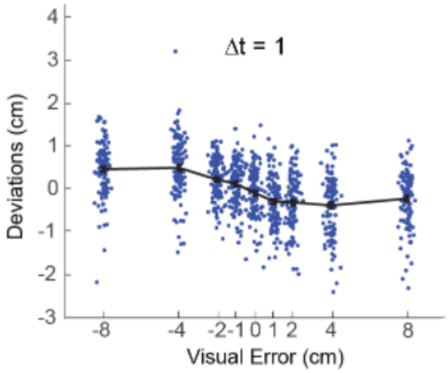
Free parameters:

- *s* is a scaling factor about the percentage of x_{vis} is corrected (or learned)
- *σ* indicates a size threshold beyond which a visual perturbation is not corrected
- *c* is a scalar, not related to x_{vis}

Importantly, we use 1) *Gaussian* noise and 2) *MLE* for the fitting of this *nonlinear* model!

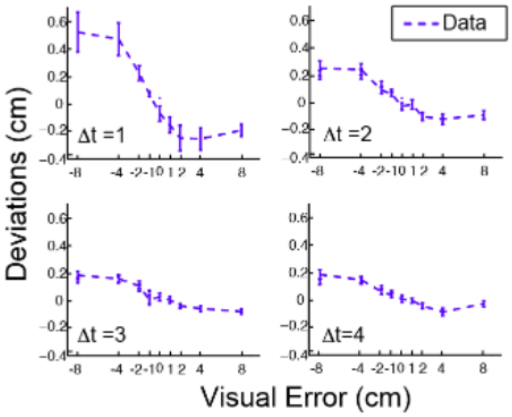
Take a look at human data first

One typical subject



Though noisy, hand deviates in the opposite direction of the visual error.

Average over subjects



- Indeed, how much people adapt is a function of visual error.
- The visual error has a decreasing influence over later trials (Δt=1~4).

Take a look at model performance

x_{hand}, x_{vis} — Actual data

$\downarrow$

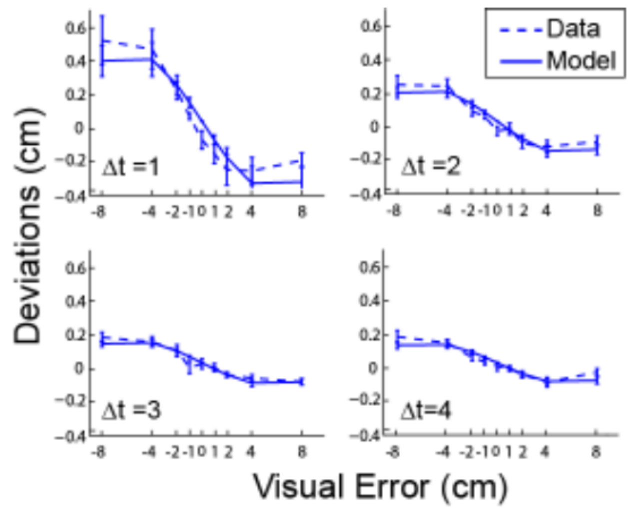
s, σ, c — Model fitting by MLE

$\downarrow$

$\hat{x}_{hand} = f(x_{vis}, s, \sigma, c)$ — Model simulation

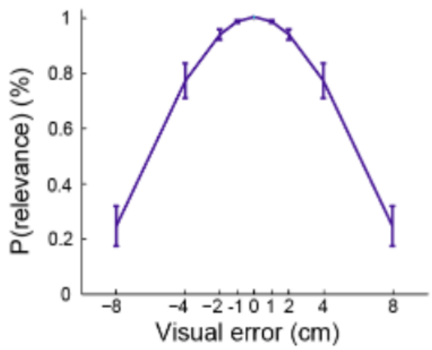
$\downarrow$

$\hat{x}_{hand}$ vs x_{hand} — Model vs data



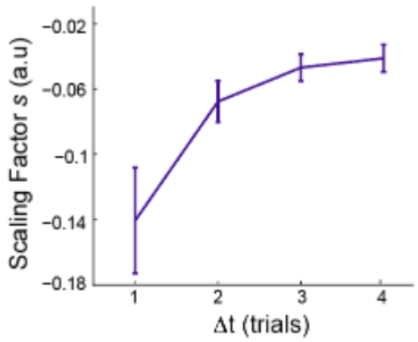
Interpret model parameters

1) Estimate the subjective belief of error relevance, as a function of visual error.



The larger the error, the less likely of a self-produced error.

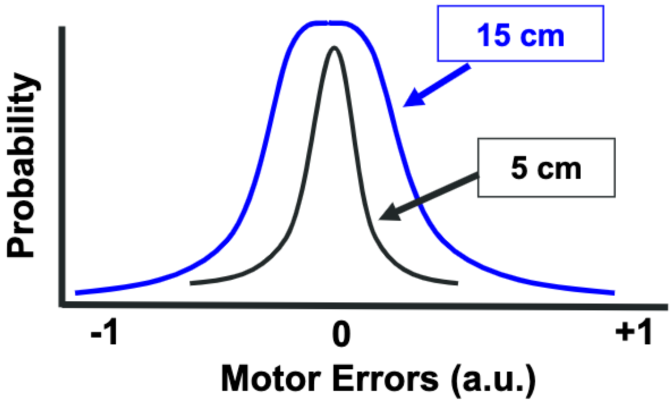
2) How much people learned from Δt previous trials?

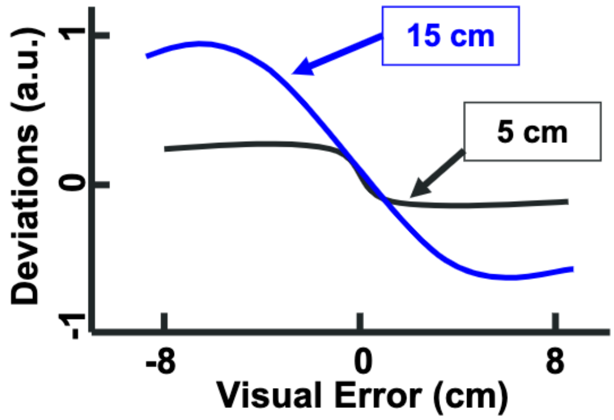


Decreasing influence of previous trials

The model makes further predictions

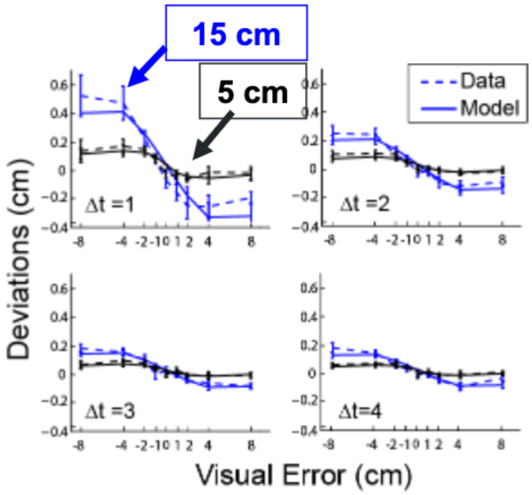
If you move shorter, then less motor noise, then less uncertainty about the error’s source, and less learning.



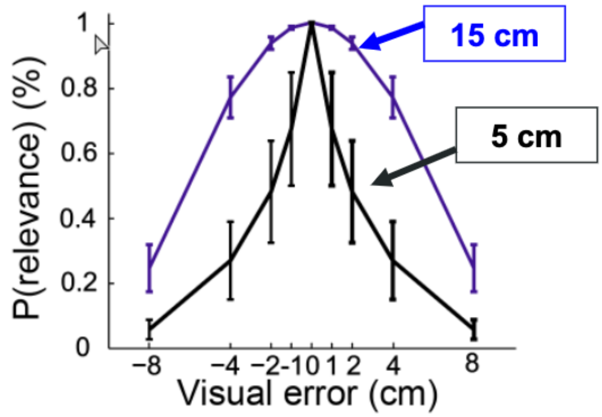


Further predictions tested in Exp2

This is not crossvalidation!



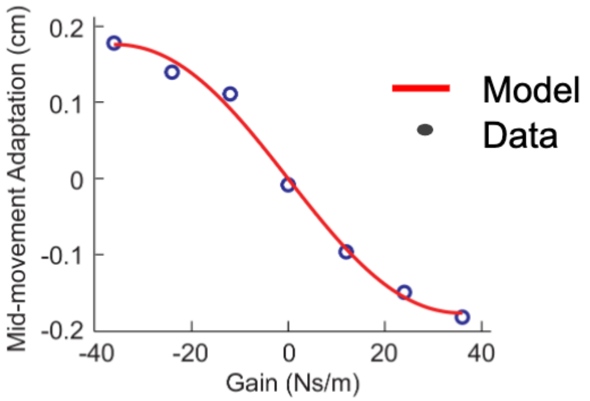
Indeed, when move shorter, less error corrections.



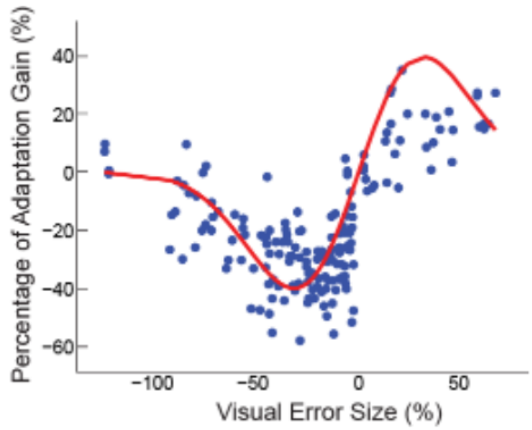
When move shorter, people have less uncertainty

More “predictions” from the model

Fitting other studies’ data, not crossvalidation here either



Adaptation to a force field (Fine and Thoroughman, 2006)



Saccadic adaptation in monkeys (Robinson et al., 2003)

Take-home messages

- Parameter estimation: Using MLE (or other methods) to estimate the model parameters, then interpret the parameters with respect to the data.
- Simulation: use estimated parameters to run simulations, check how good the fits are, and make new predictions. Predictions can be out of the parameter space of your original dataset.

We strive for:

- Linking neural properties, computational mechanisms and behavior.
- Understand the mechanisms.
- New predictions.

(3) Optimization and convex problems

Linear models

and

Generalized linear models (GLM)

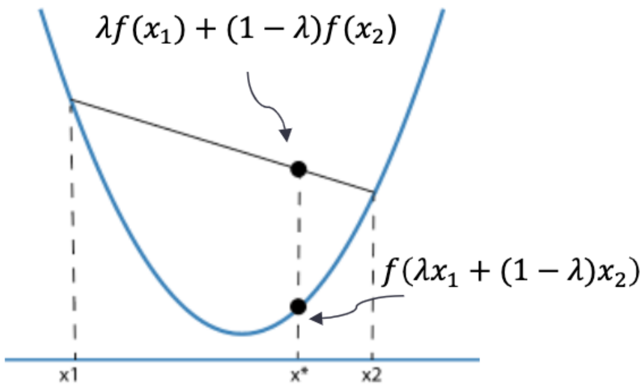
MLE for linear models and convexity

- Quadratic likelihood function for linear models:

$$\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y}) = \sum_n \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}_n, y_n)$$
$$= -\frac{1}{2\sigma^2}\sum_n \boxed{\left(y_n - \boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\right)^2} + \text{const.}$$

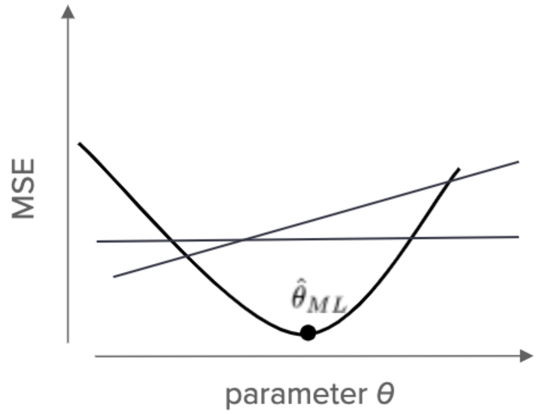
- Jensen‘s inequality for convexity:

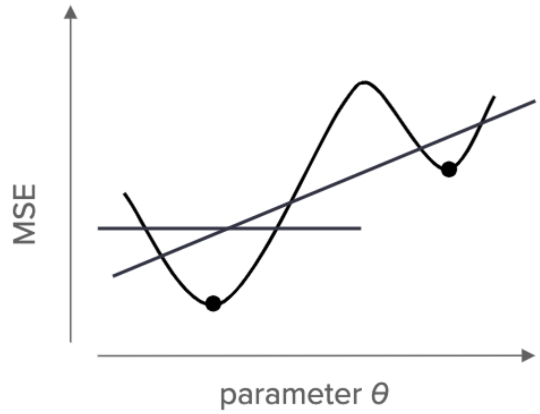
$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$



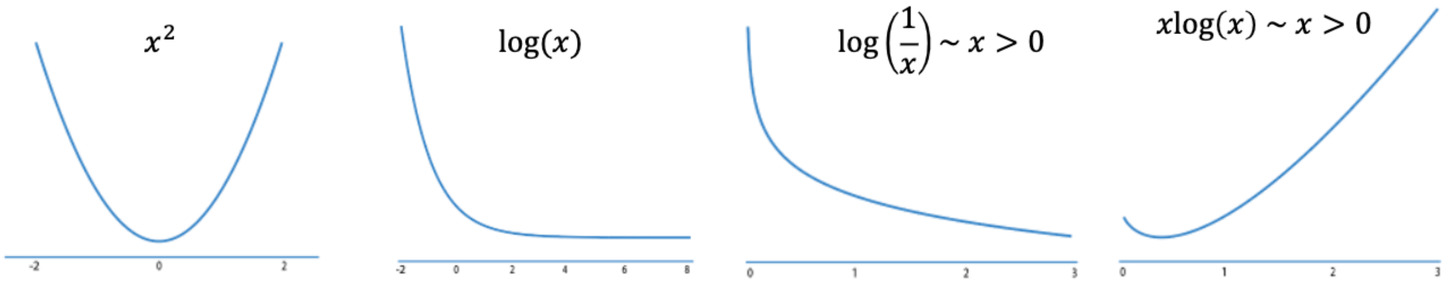
- Quadratic function is convex with a single minimum.

Convexity prevents local minima





Convex exponential functions



Generalized linear model (GLM)

- Linear model

$$y = \theta x + \eta$$

$$\eta \sim N(0, \sigma^2)$$

- Generalized linear model

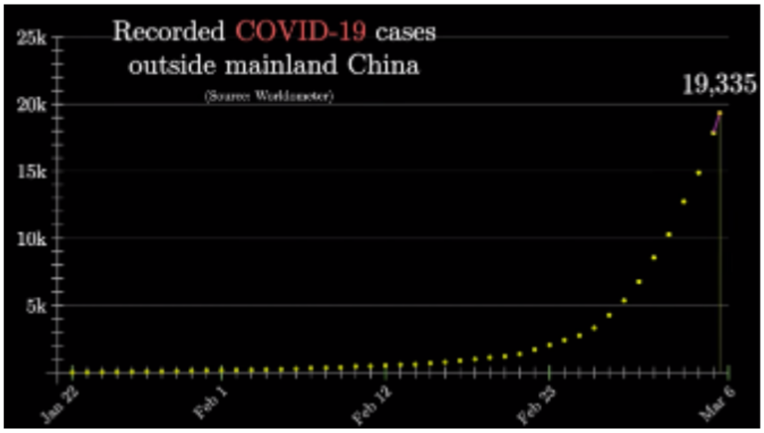
$$y = f(\theta x) + \eta \qquad \eta \sim N(0, \sigma^2)$$

$$y \sim Distribution(f(\theta x))$$

where θx is still the linear predictor. However, the link function f^{-1} allows *y* follow **other exponential distributions**, including binomial, Poisson, gamma distributions and so on.

- GLM covers a broad set of models!

One GLM example: exponential



Source: 3Blue1Brown

- The number of COVID-19 case y_i follows an exponential function of time t_i :

$$y_i = \gamma e^{\delta t_i}$$

- As a GLM model, a link function is log()

$$\log y_i = \log(\gamma) + \delta t_i = \theta_0 + \theta_1 x_i$$

- Now it is a familiar linear function

The nice thing: GLM is convex

- The GLM log-likelihood is concave in the parameters
- **Thus, finding the minimum is relatively easy for GLM models**
- **Details of GLM will be covered in Day 4.**