Regression & Linear Models

BIPN 162

Pre-Class Viewing

$$y \approx \theta_0 + \theta_1 x$$

intercept slope

y = mx + b

You should be able to:

- List four different reasons for fitting models to our data
- Write & interpret the generic form of linear model
- Describe two different philosophies on model fitting:
  - minimizing error **(descriptive)**
  - maximizing likelihood **(generative)**
- Describe how **bootstrapping** can be used to assess uncertainty
- Explain the trade-off between underfitting & overfitting your data

Objectives for today

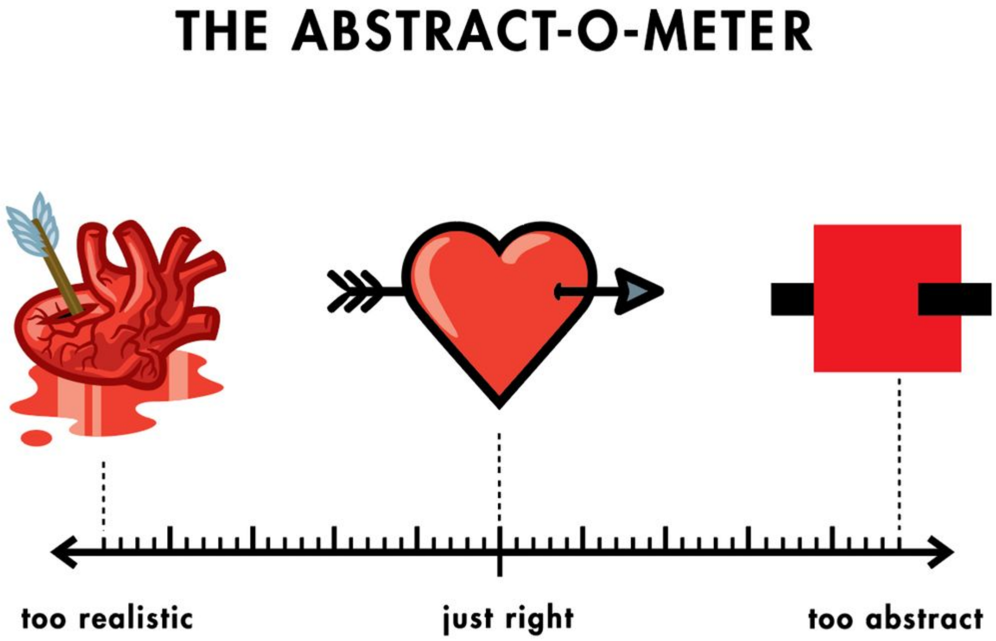
- **Motivate the use of models to describe neural data**
- Recap what we've discussed about distributions & correlations
- Identify multiple ways to assess the fit of a model
  - Define **mean squared error** and **maximum likelihood estimation**
  - Describe the process of **bootstrapping**
- Implement a linear regression for **data exploration** and **prediction**

What are models?

Models are an abstraction of reality!

How do we find the right level of abstraction?

- Keep it as simple as possible, but as detailed as needed
- This is determined by our question, hypotheses and model goals!



Slide adapted from Neuromatch Academy
Image from https://computersciencewiki.org/index.php/Abstraction

What are models?

Models allow for **understanding** and **control** (Rosenblueth & Wiener, 1945)

Insights not directly accessible by experiments / data

Interventions, e.g. experimental, clinical

Requires model validation → experiments!

In other words, a model is a **Hypothesis**!

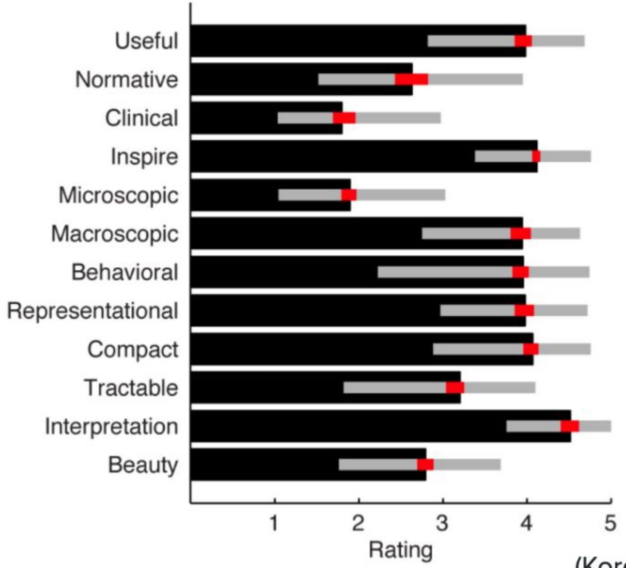
A mathematical one!

Sometimes we even *compare* models: <u>model comparison</u>

Model diversity

- Different models allow answering different questions
- Everyone has a different idea of what's a "good" model

- "A good model" is what best answers the question with minimal assumptions and highest explanatory power
- Thus **model diversity** is good!



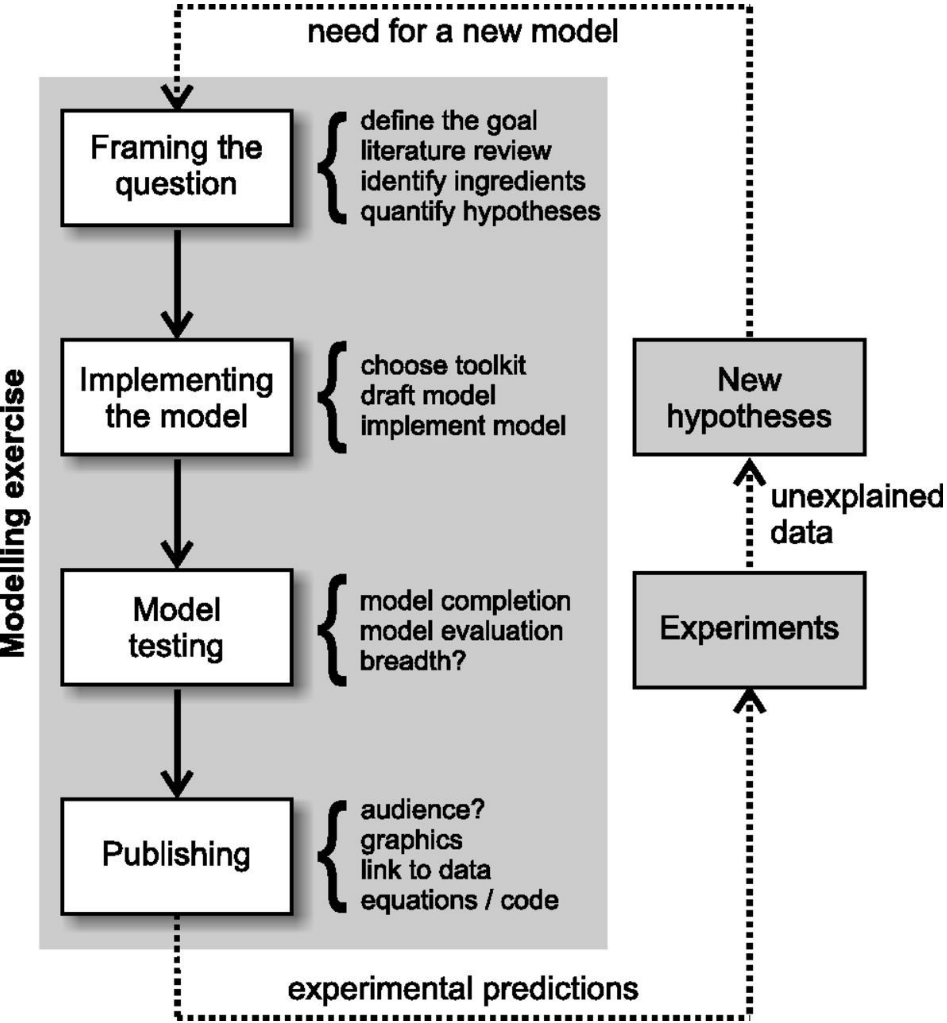
(Kording, Blohm, Shrater, & Kay https://osf.io/3vy69/)

Adapted from NMA; figure from https://osf.io/3vy69/

Asking good questions that we can model

not great	better!
How does a clock work?	How do the angles of the hands of a clock predict the time of the day?
Can we predict whether a rat will turn left or right from neural activity?	Which neural activity, when, and which types of analyses significantly predict whether a rat will turn left or right?
How does the motor system work?	How does sensory noise influence motor movements in this experiment, and why?

Adapted from NMA



The modeling process

(from [A How-to-Model Guide for Neuroscience | eNeuro])

Regression is a method that allows researchers to summarize how predictions or average values of an outcome vary across individuals defined by a set of predictors

it can be used for....

- prediction
- exploring associations
- causal inference *(more on this later!)*

Adapted from [Regression and Other Stories](), Chapter 1

Linear regression makes predictions about the linear relationship between input variable(s) & output variable

"**coefficient**", "**parameter**", unknown

dependent
known/observed

$$y = \beta x$$
$$y = x_1\beta_1 + x_2\beta_2$$

independent
known/observed; "**predictor**"

For linear regressions...

$$Y = X\beta + \beta_0 + \epsilon_1 + \epsilon_2 + \dots$$

error source 1 error source 2 ...

Slide adapted from Marcus Benna

Linear combinations and the CLT

$$Y = X\beta + \beta_0 + \epsilon$$ ← **error / noise**

where $\epsilon = \epsilon_1 + \epsilon_2 + \ldots$

$$\boxed{\epsilon \xrightarrow{d} \sim \mathcal{N}(0, \sigma^2)}$$

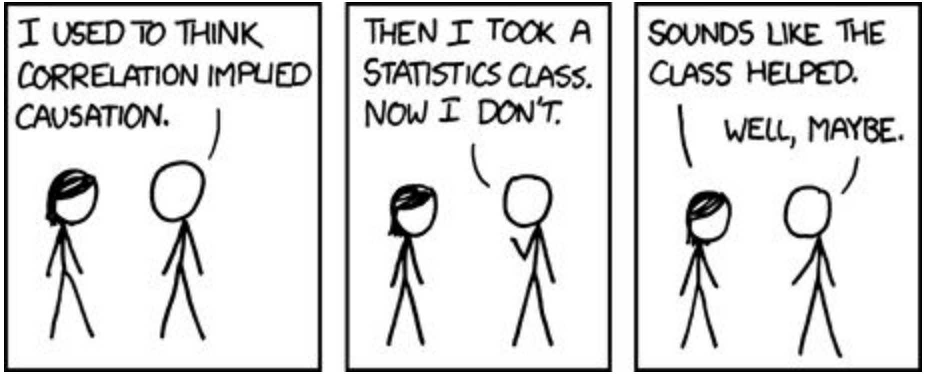
"Convergence of distribution", in other words, error is normally distributed

$$Y \sim \mathcal{N}(X\beta + \beta_0, \sigma^2)$$

Linear regression with Gaussian errors

Slide adapted from Marcus Benna

Objectives for today



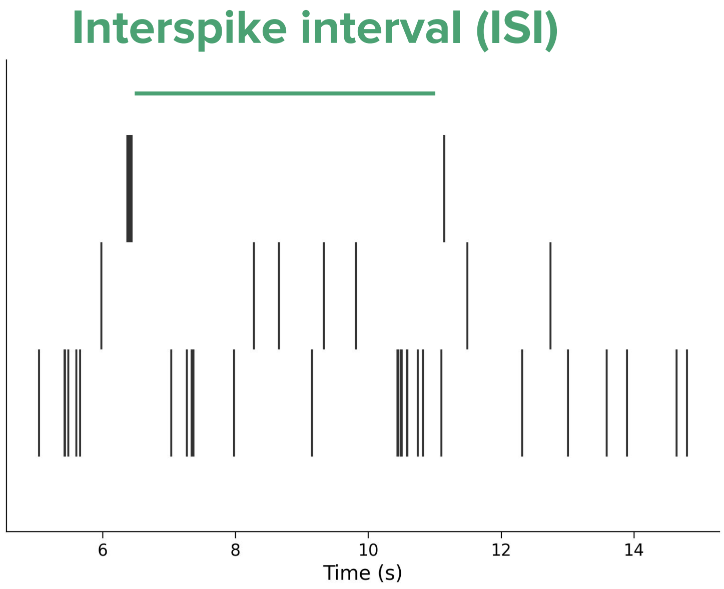
https://xkcd.com/552/

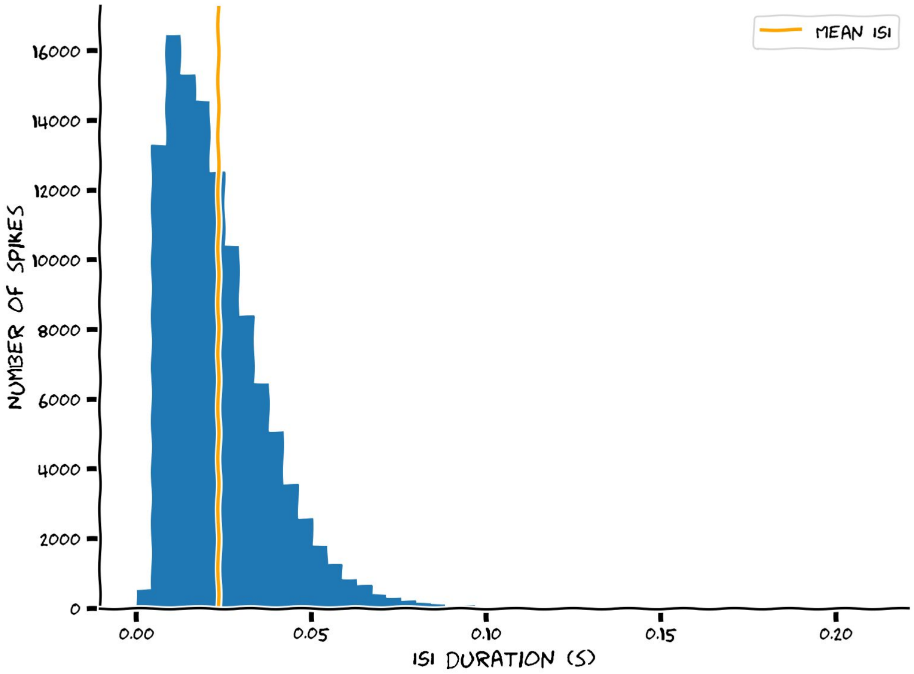
- Motivate the use of models to describe neural data
- **Recap what we've discussed about distributions & correlations**
- Identify multiple ways to assess the fit of a model
    - Define **mean squared error** and **maximum likelihood estimation**
    - Describe the process of **bootstrapping**
- Implement a linear regression for **data exploration** and **prediction**

We’ve already talked about models!

When we build distributions, we are building a **model** of what we think the data looks like

As another example of this, we could look at the interspike intervals (ISIs) between neurons and ask what kind of distribution they fit





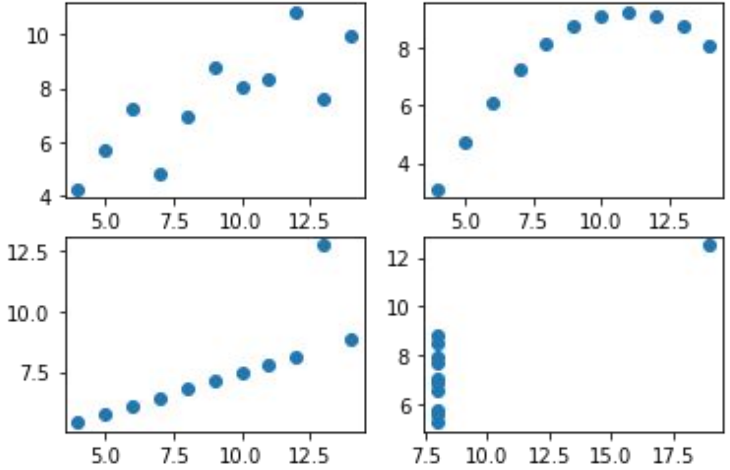
What distribution would best fit the decay of the ISI?

(In other words, what equation/function can we use to model it?)

We've also discussed **correlation**

The Pearson correlation coefficient can be computed by:

1. Mean centering each variable
2. Dividing **dot product** of mean centered variables by **product** of mean centered variables



Reminder: Pearson correlation only measures **linear association**

What's the difference between r and R^2?

r	$\mathbf{R^2}$
correlation coefficient (ρ, rho)	coefficient of multiple determination
measure of the strength of the **linear** relationship between two variables	measure of the strength of the **regression** equation, or how well we can predict *y*
ranges from -1 to +1	ranges from 0 to +1
Differences are less intuitive	Differences are more intuitive, in other words, $R^2 = 0.6$ is twice of good as a fit as $R^2 = 0.3$

Additional explanation: https://www.youtube.com/watch?v=2AQKmw14mHM
Play with r values: https://www.esci.thenewstatistics.com/esci-correlation.html

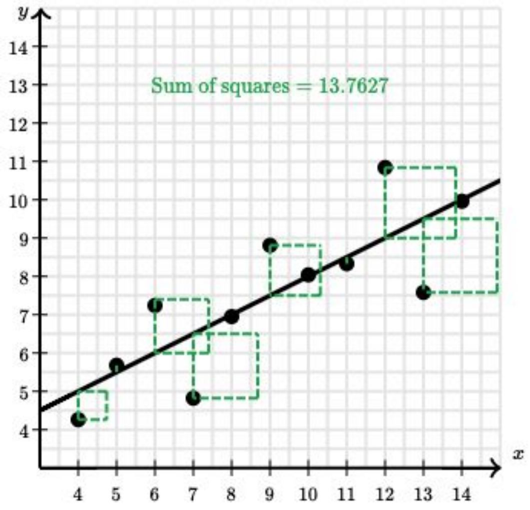
Objectives for today

- Motivate the use of models to describe neural data
- Recap what we've discussed about distributions & correlations
- **Identify multiple ways to assess the fit of a model**
  - Define **mean squared error** and **maximum likelihood estimation**
  - Describe the process of **bootstrapping**
- Implement a linear regression for **data exploration** and **prediction**

How do we assess our fit?

Descriptive approach

- Previously, we described (ordinary) **least-squares regression** to find a line of best fit and derive an R^2 value
- We can also use **mean squared error (MSE)**, where we *average* the square differences between data and prediction (the **residuals**)

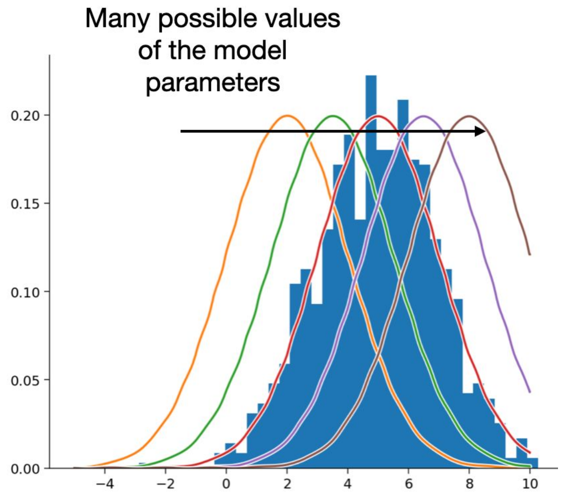


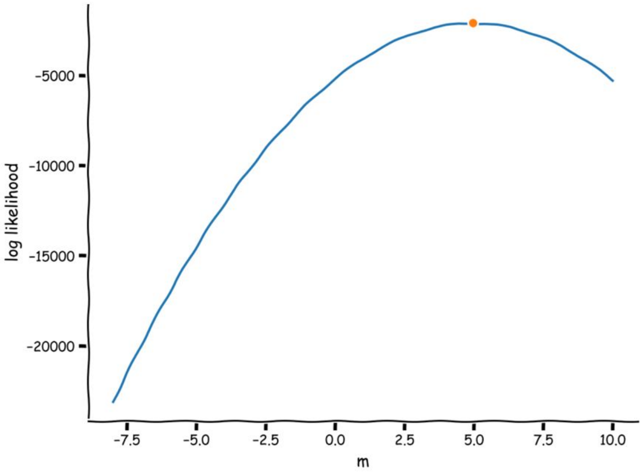
$$\mathrm{MSE} = \overbrace{\boxed{\frac{1}{n}\sum_{i=1}^{n}}}^{\text{Mean}} \Big(\overbrace{\boxed{Y_i - \hat{Y}_i}}^{\text{Error}}\Big)^{\overbrace{\boxed{2}}^{\text{Squared}}}$$

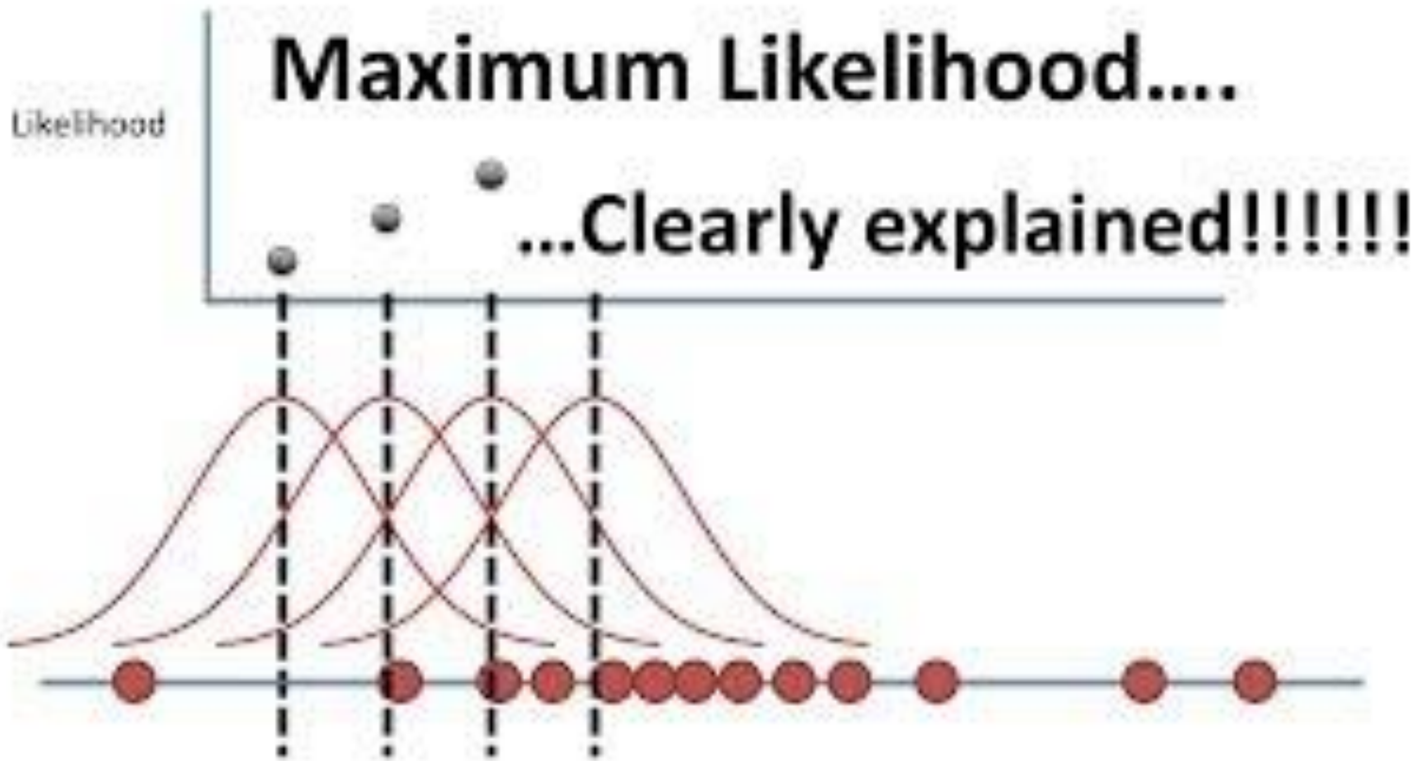
How do we assess our fit?

Generative approach: fitting models by **maximum likelihood**

- What are the parameters that make the data most likely?
- If we maximize the likelihood with gaussian noise, **this is equivalent to minimizing the MSE!**
- Usually calculated as **log** likelihood





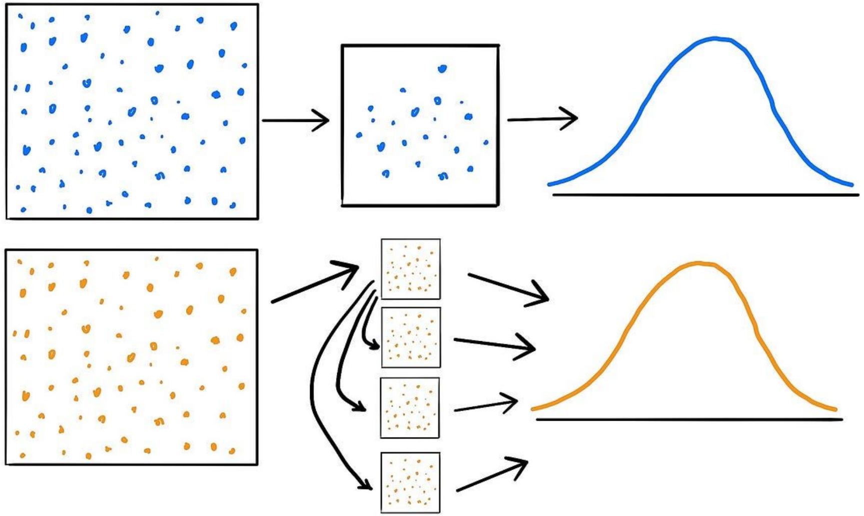


https://youtu.be/XepXtl9YKwc?si=m75a9gjOJSgjQsOg; see also
https://compneuro.neuromatch.io/tutorials/W1D2_ModelFitting/student/W1D2_Tutorial2.html

We can also use **bootstrapping** to assess confidence/uncertainty about estimated parameters

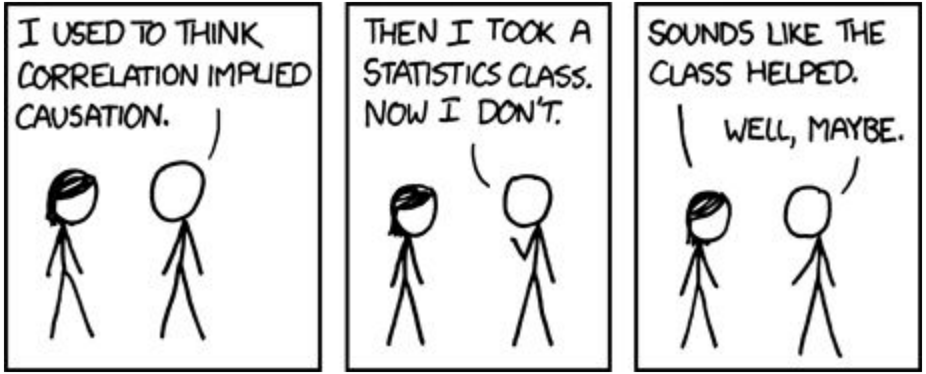
Steps:

1. Generate many new *synthetic* datasets from the initial true dataset by **randomly sampling** from it (*with replacement*)
2. Find estimators for each one of these new datasets
3. Inspect the distribution of all these estimators to quantify our confidence



From What Is Bootstrapping Statistics?

Objectives for today



https://xkcd.com/552/

- Motivate the use of models to describe neural data
- **Recap what we've discussed about distributions & correlations**
- Identify multiple ways to assess the fit of a model
  - Define **mean squared error** and **maximum likelihood estimation**
  - Describe the process of **bootstrapping**
- Implement a linear regression for **data exploration** and **prediction**

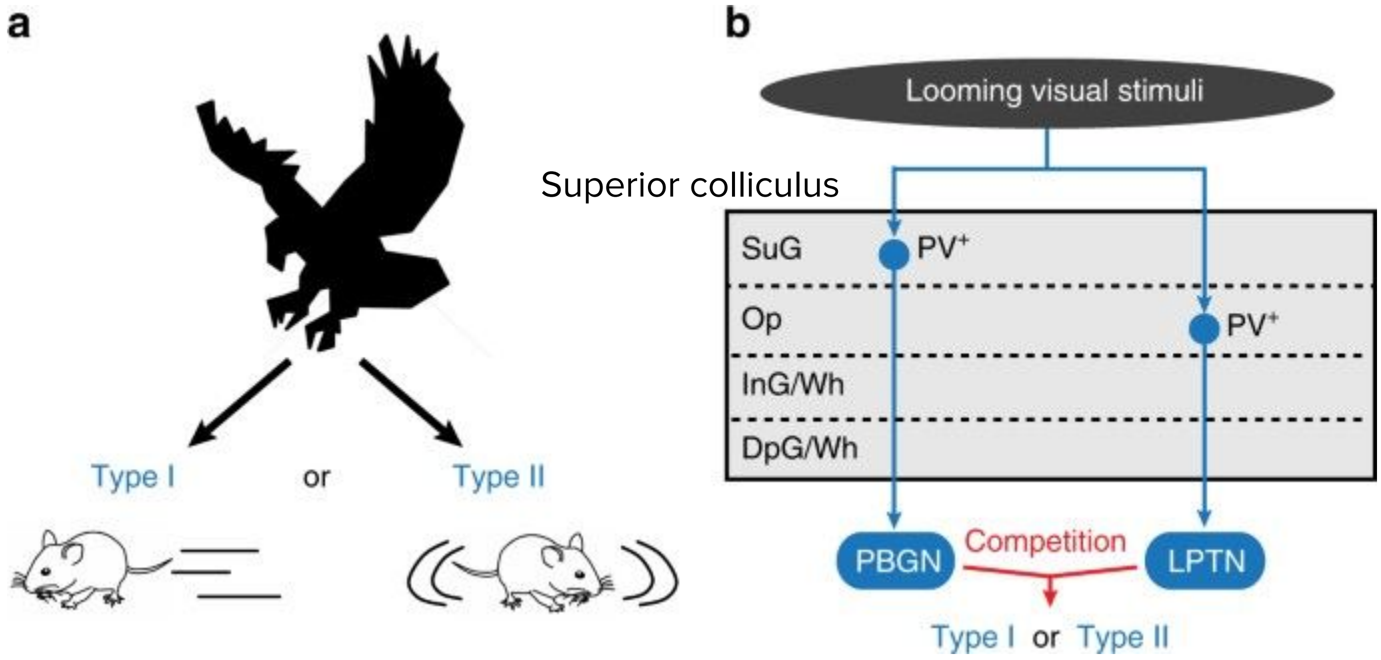
Can we predict behavior using a linear model?

The simplest movement for a mouse?
Running!

- Movement drives activity across the cortex (Stringer et al., 2019, Musall et al., 2020)
- Running de-correlates neural activity — it reduces pairwise correlations (Erisken et al., 2014)
- Running sharpens tuning of primary visual cortex (V1) neurons (Neill & Stryker, 2013)
- V1 integrates running speeds to compute visual motion (Saleem et al., 2013)

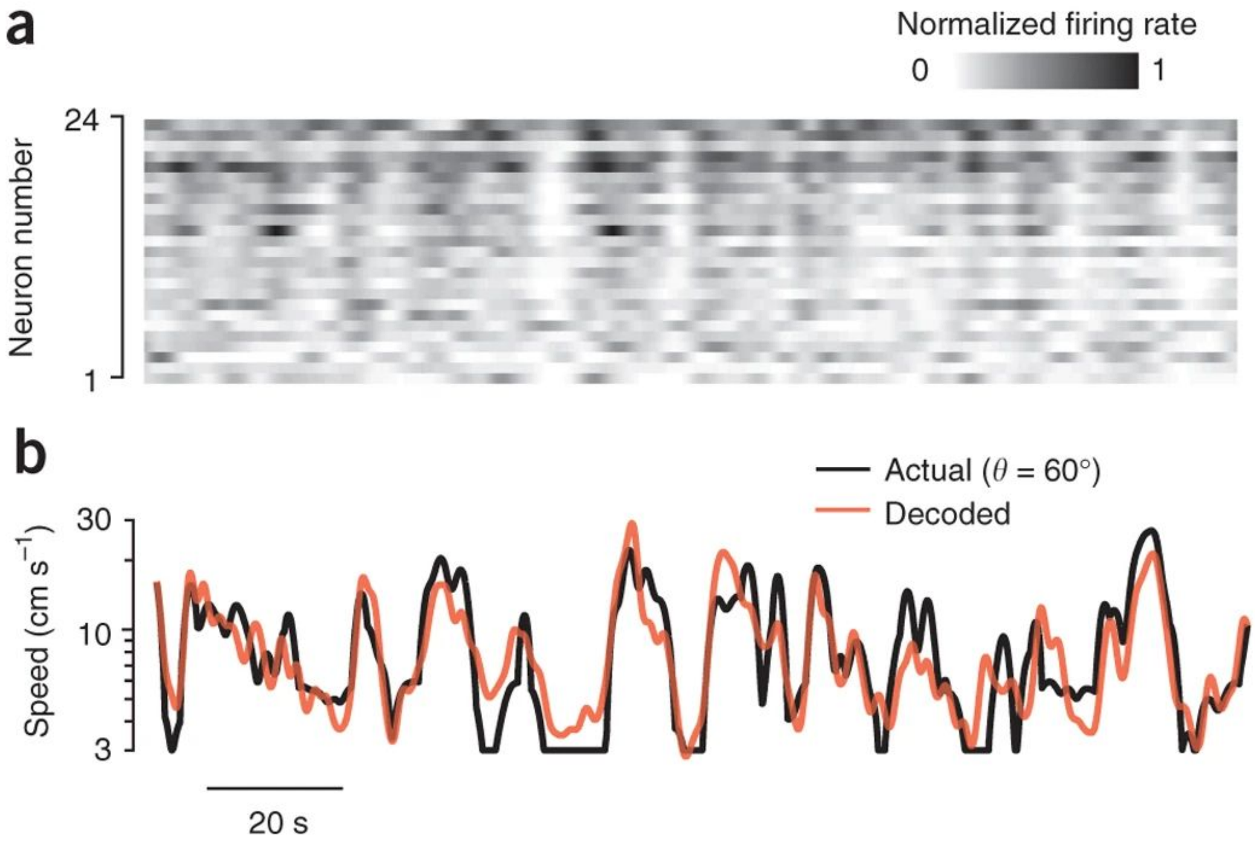
Image source

Running has an important ethological benefit for mice...



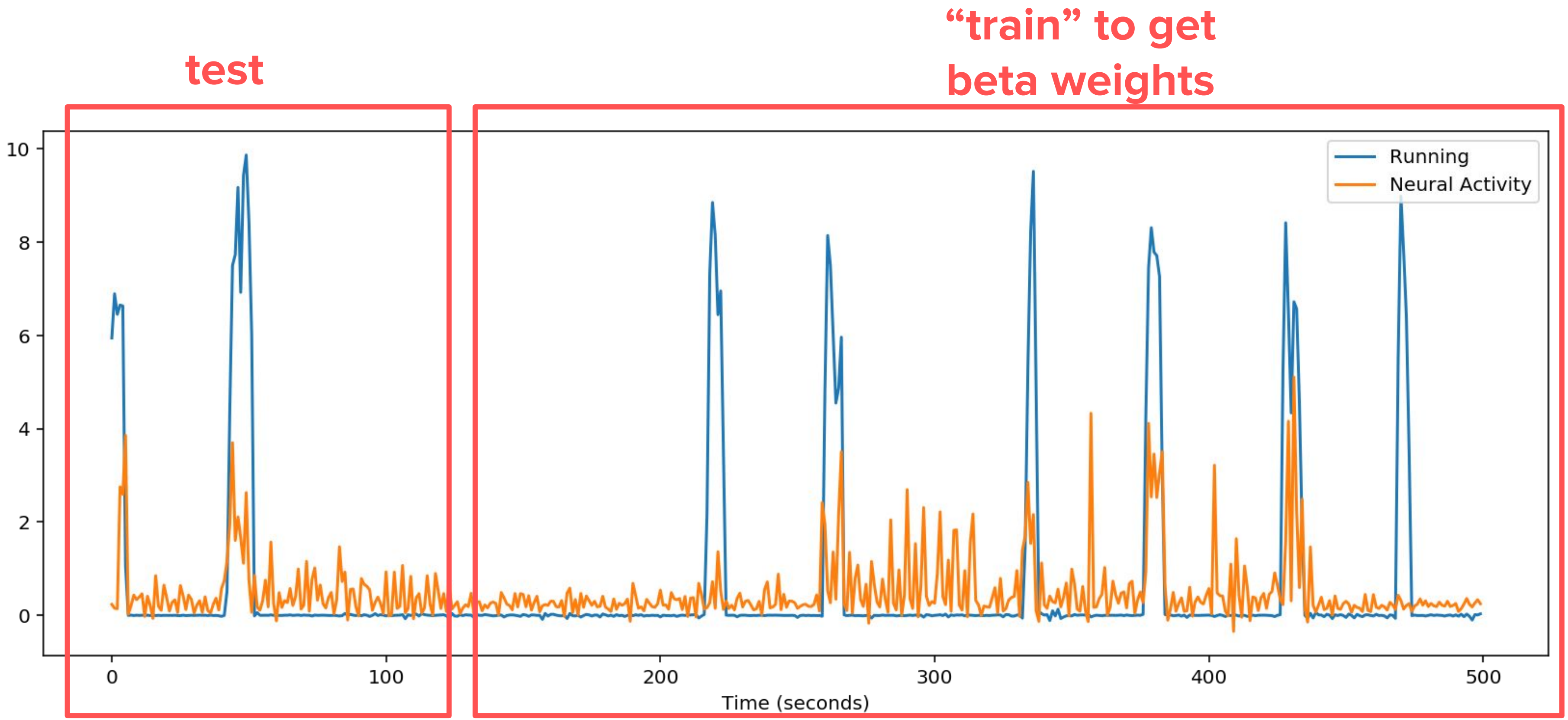
Circuit dissection of escape behaviors by [Shang et al., 2018](#)

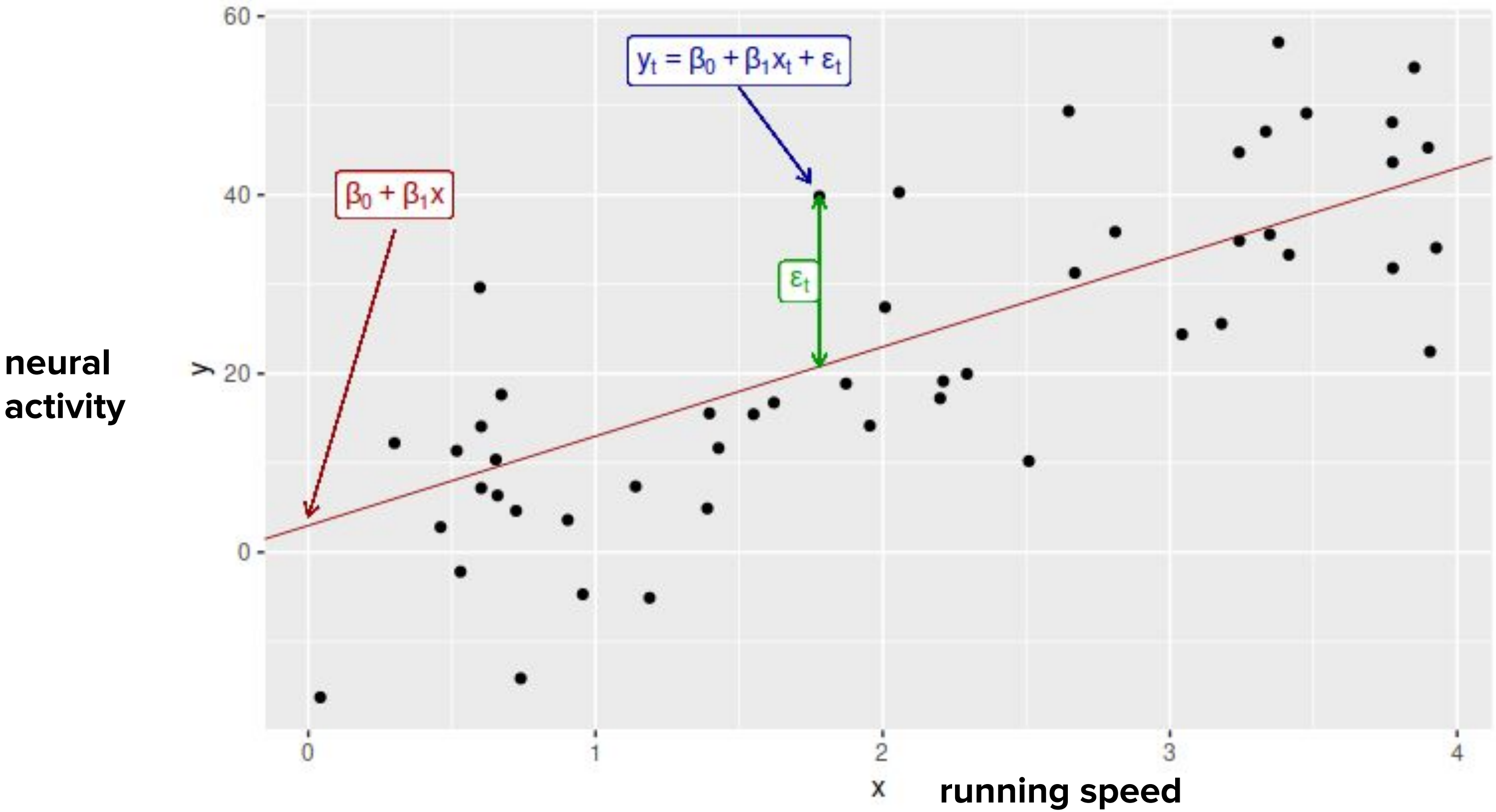

Locomotion similar to hindlimb unloading



Even with a linear model, we can decode running speed (from Saleem et al., 2013)

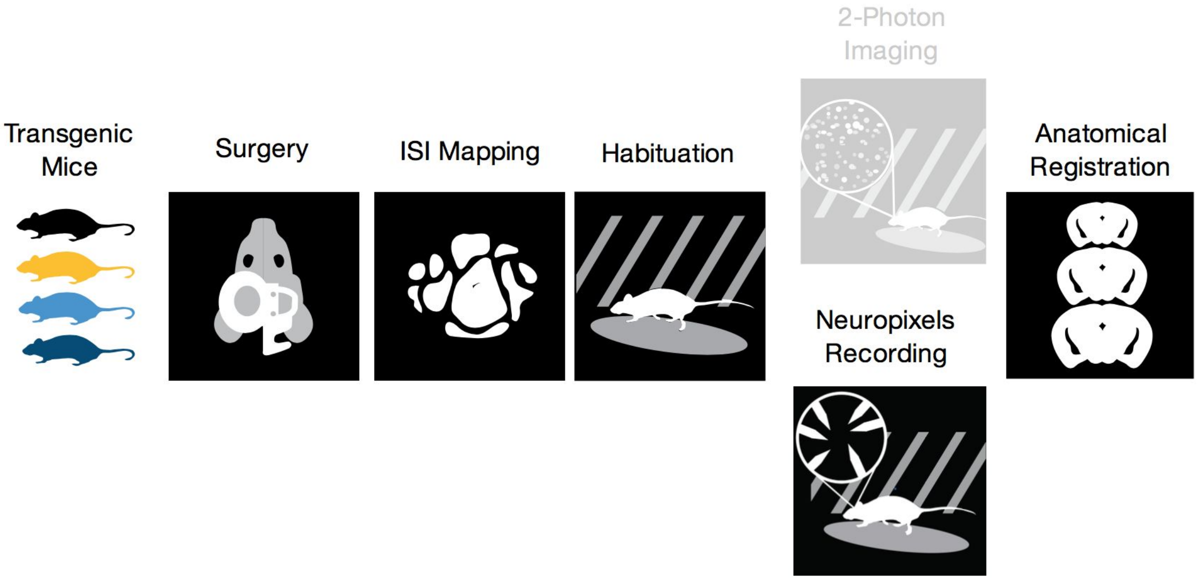
How do we build a predictive linear regression model?





The coefficients **β0** and **β1** denote the intercept and the slope of the line, respectively. The intercept β0 represents the predicted value of y when x = 0. The slope β1 represents the average predicted change in y resulting from a one unit increase in x. [source]

Both the Brain Observatory & Neuropixels datasets have information about the animal's running speed and pupil diameter.



(Ordinary) **least squares regression**

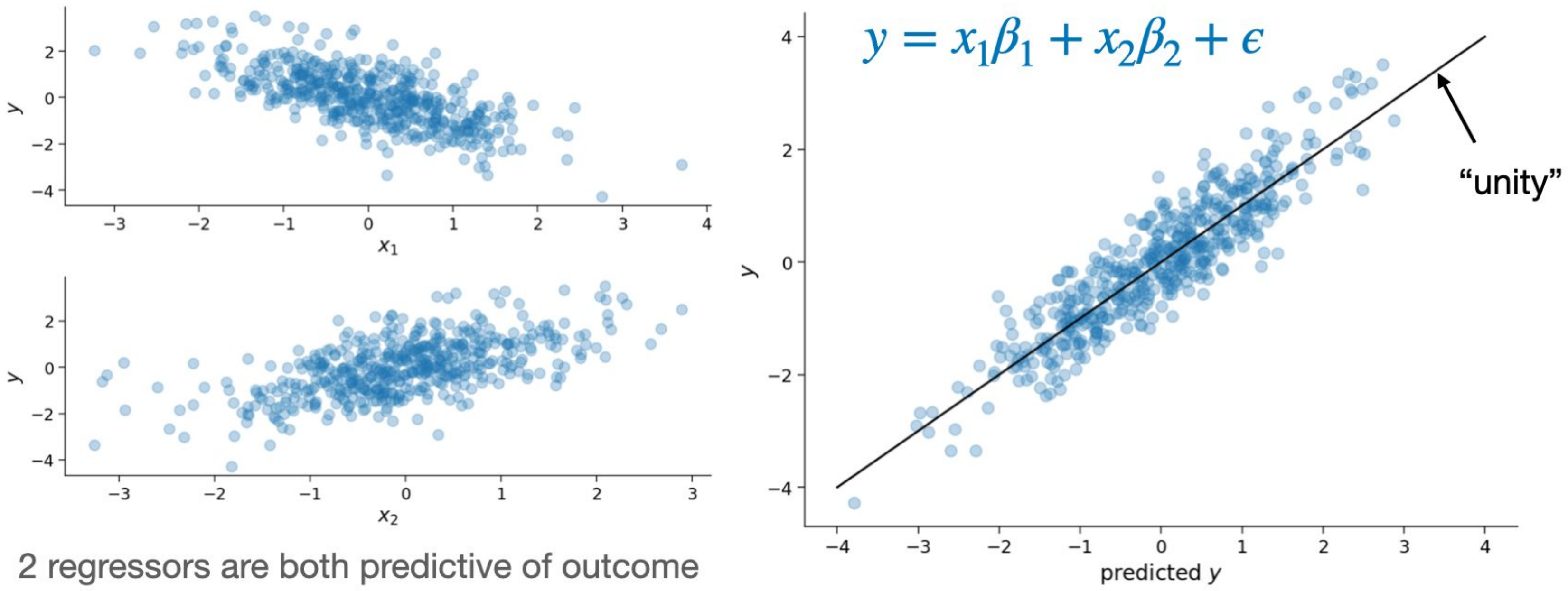
- We'll use a linear regression model to see if we can predict neural responses based on running speed.
- We will implement this using **`sklearn.linear_models.LinearRegression()`**
- We will assess our model by the R^2 value
- We can also compute the **mean squared error** — the regression line is the line that minimizes this value

Regression models with more than one predictor/regressor

In neuroscience, we often have more than one predictor variable:

E.g., in a clinical study, we might like to know how age, gender, and years playing soccer predict lifetime head trauma.

Regression models with more than one predictor/regressor



2 regressors are both predictive of outcome

Slide adapted from Mikio Aoi/Marcus Benna

Gaussian, Bernoulli and Poisson distributions are all members of wider class of distributions, known as the **exponential family.**

We can perform a regression by modeling the response Y as coming from a particular member of this family, and then **transforming the mean** of the response so that the transformed mean is a linear function of the predictors via

$$\eta(\mathrm{E}(Y|X_1,\ldots,X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Any regression approach that follows this very general recipe is known as a **generalized linear model (GLM)**.

Resources

An Introduction to Statistical Learning !!!!

Tutorial 1: Linear regression with MSE — Neuromatch Academy: Computational Neuroscience

Pearson Correlation and Linear Regression

Correlation and linear regression