Linear models

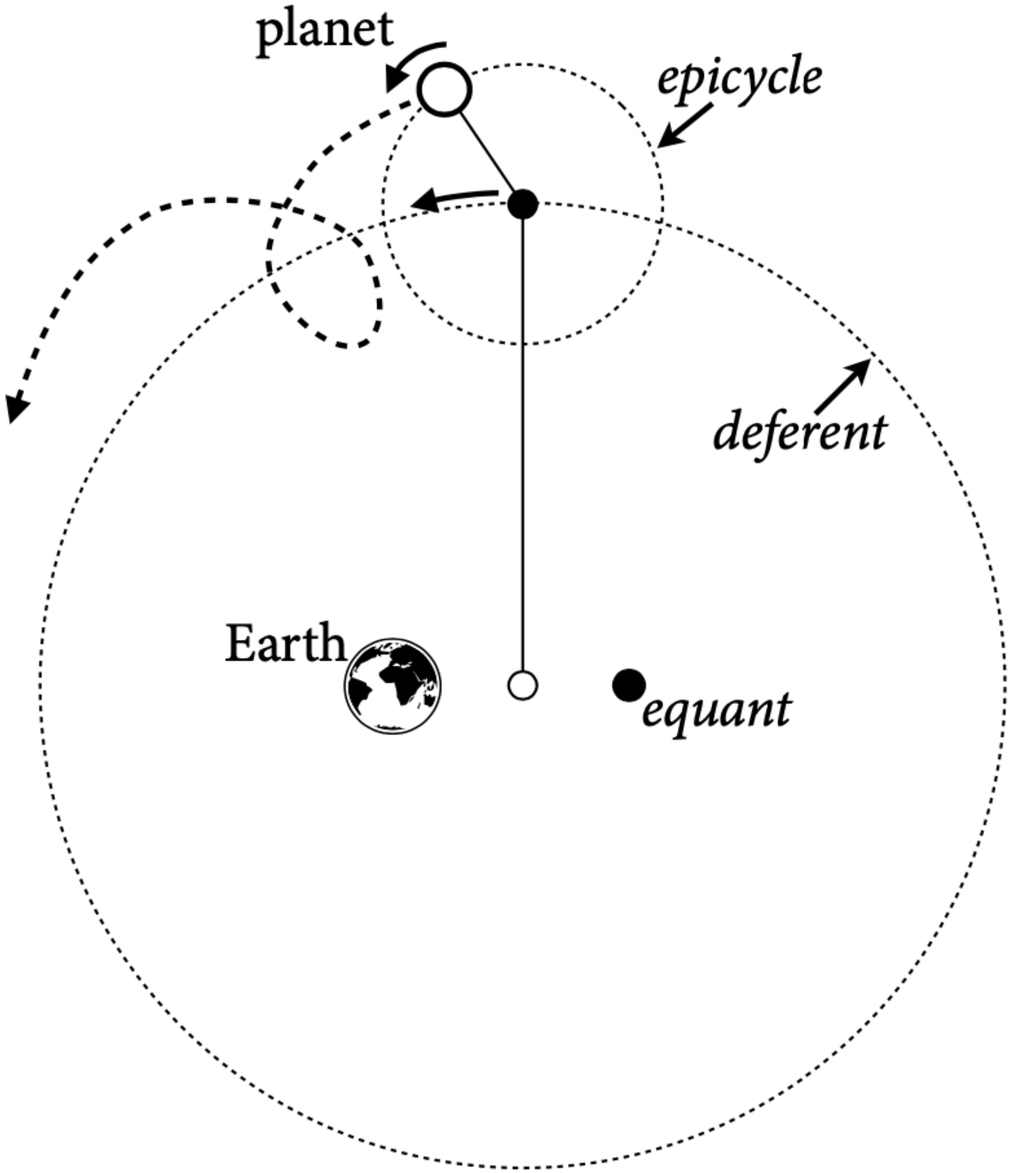
Data analytics

Jerzy Baranowski

Geocentric models

Lets add some epicycles

- Ptolemeic model of the solar system
- Very accurate - causally wrong.
- Many statistical models are very good at predictions, but that not necessarily mean that there are causal relationships.



Regression

Or what is with nomenclature

- Generally regression models are used to explain relationships between independent and dependent variables
- Practically it means that we create models that are used to determine a numerical value

Why Gaussians are so popular?

Or a bit of back to basics

- Notation

$$y \sim \text{Normal}(\mu, \sigma)$$

is understood as y is a random variable with normal distribution with a mean μ and standard deviation σ.

- In Bayesian thinking it means, is that we are uncertain about value of y but our uncertainty can be described by Normal distribution.

So what are statistical models?

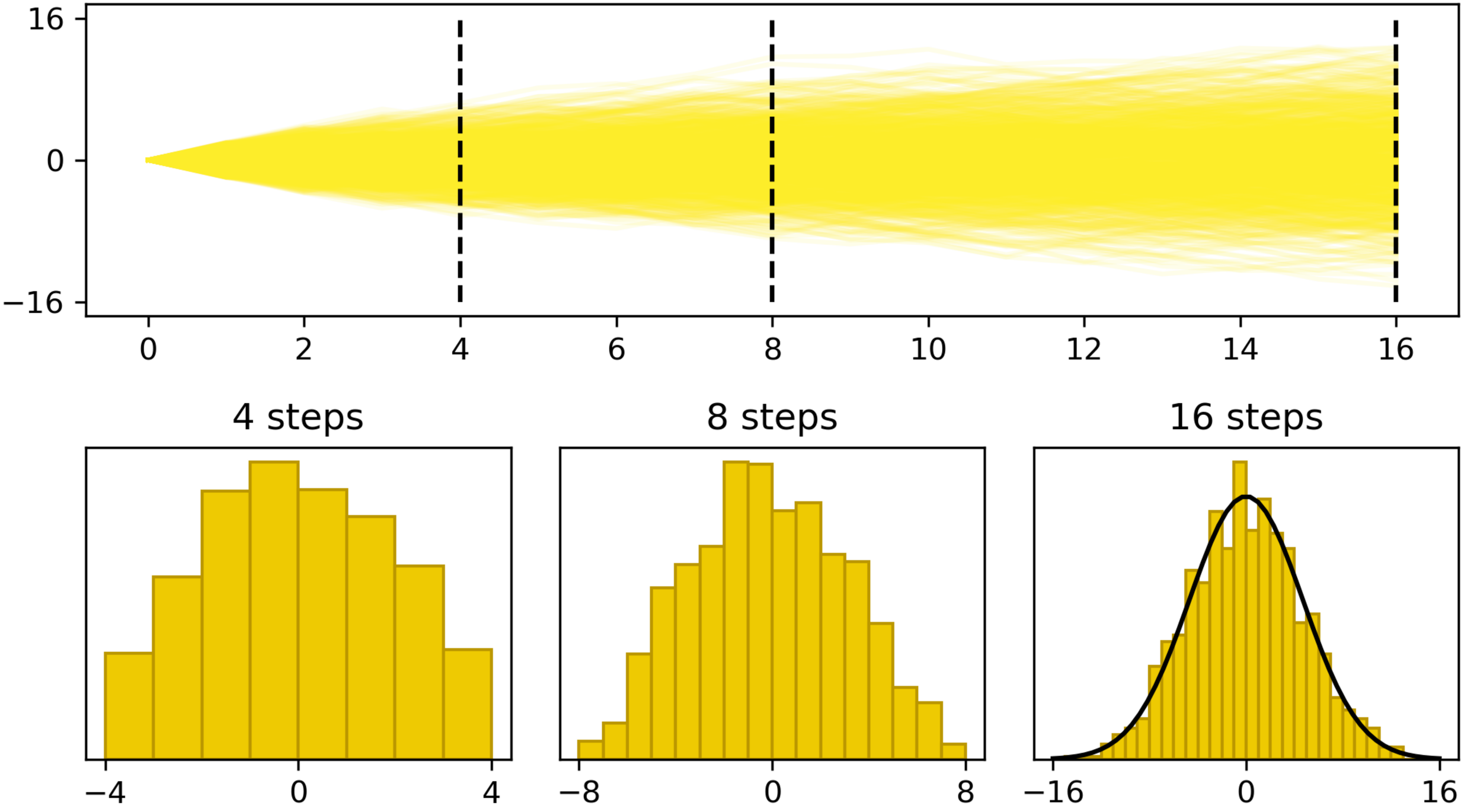
And what to do with them

- We want to find the description of our uncertainty about a problem in order to be able to make predictions or make inferences about that problem.
- Statistical models are like robots - they do the given task processing the data that is available to them. If we give them rubbish we will get rubbish.
- In Bayesian frameworks we get models in the form of probability distributions. That encapsulates our uncertainty and we can use that distribution to simulate possible outcomes representing our uncertainty.

But why gaussians?

Spoiler, they are everywhere

- Gaussian distribution arises from addition.
- Central limit theorem says, that sum of iid random variables is normally distributed
- Thought experiment:
  - Imagine 1000 people, we put them in the middle of football field, and tell them: „You need to make 16 steps along the white line in any direction you want. You can change direction as you wish. Same with step length”. What is the distribution of positions?
  - For convenience we assume, that steps are distributed as $\mathrm{Uniform}(-1,1)$


16
0
−16
0
2
4
6
8
10
12
14
16
4 steps
8 steps
16 steps
−4
0
4
−8
0
8
−16
0
16

So what else normal distribution is good for?

Multiple things

- Processes that come in a form of series of increments are normally distributed.
- This is also true for small percentage changes (multiplication by numbers close to 1)
- In case of multiplication by large numbers we have normality on logarithmic scale.

Construction of a model

An algorithm

1. Recognize the set of measurements that determine the outcome.
2. Define likelihood i.e. plausibility of individual observations (in linear models it is usually Gaussian)
3. Recognize the measurements that we want to use to predict the outcome.
4. Relate the likelihood distribution to the predictor variables. We define model parameters.
5. Choose priors for model parameters

How to construct such model?

Bayesian model formalism example

$$
\begin{aligned}
\text{outcome}_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta \cdot \text{predictor}_i \\
\alpha &\sim \text{Normal}(0,10) \\
\beta &\sim \text{Normal}(0,10) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

Mean is parametrically related to the predictors. This is almost classical linear regression, except that classical priors are flat.

But why linear function of parameters?

Locality of models

- Linear models are justified by Taylor series expansion

$$f(x, \theta) = f(x_0, \theta) + \frac{\mathrm{d}}{\mathrm{d}x} f(z, \theta)\Big|_{z=x_0} \cdot (x - x_0) + \ldots$$

$$= \alpha + \beta \cdot x + \ldots$$

- Any nonlinear relationship of predictors to outcome can be locally approximated by a linear function of predictors with constant parameters

Lets do some linear modeling

The tall and short story of !Kung

- !Kung are a part of the San people who live mostly on the western edge of the Kalahari desert, Ovamboland (northern Namibia and southern Angola), and Botswana. There is a well documented dataset from surveying one of such tribes (Nancy Howell, 1960s)