Lecture 7 - Linear models

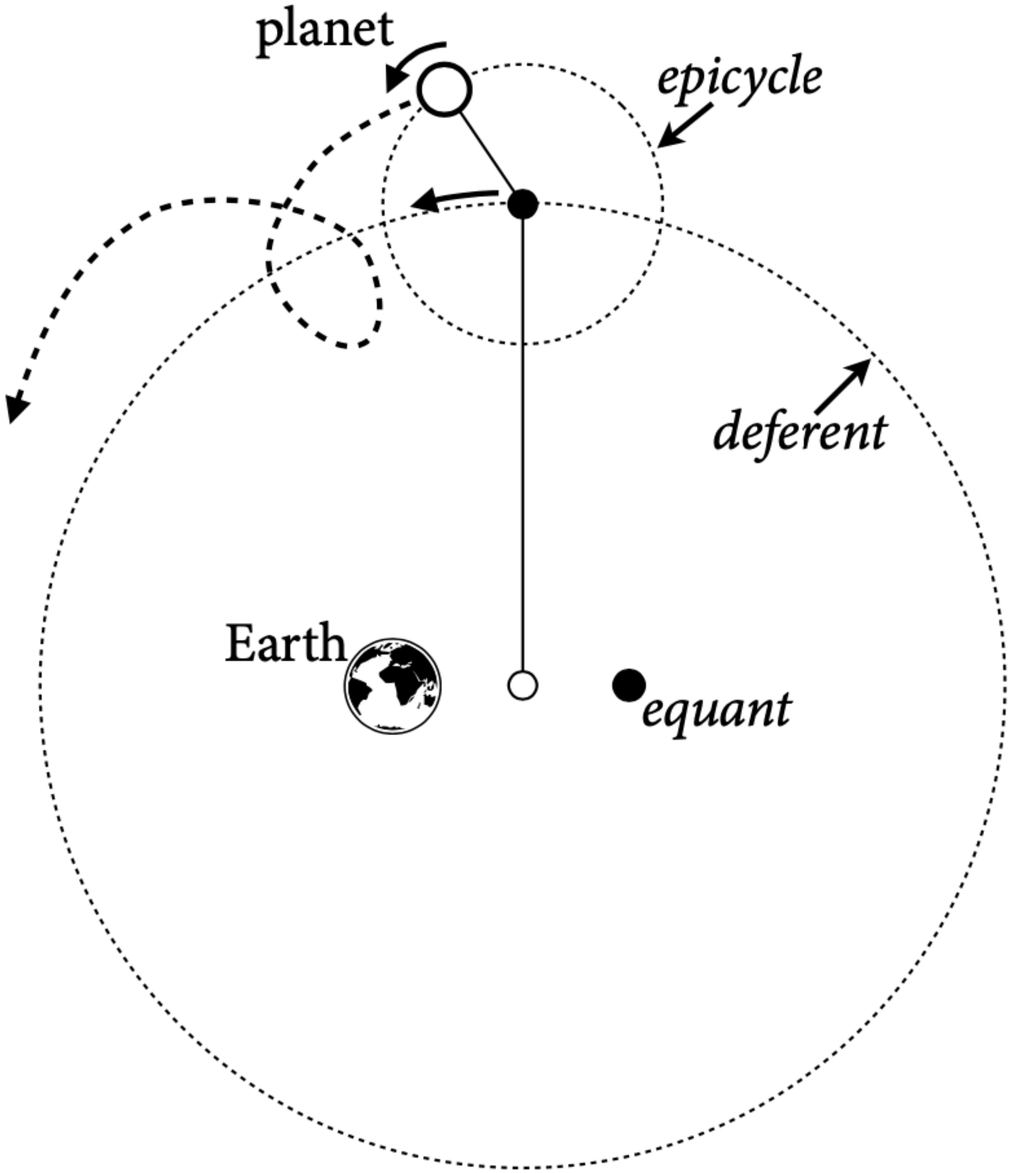
Data analytics

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?

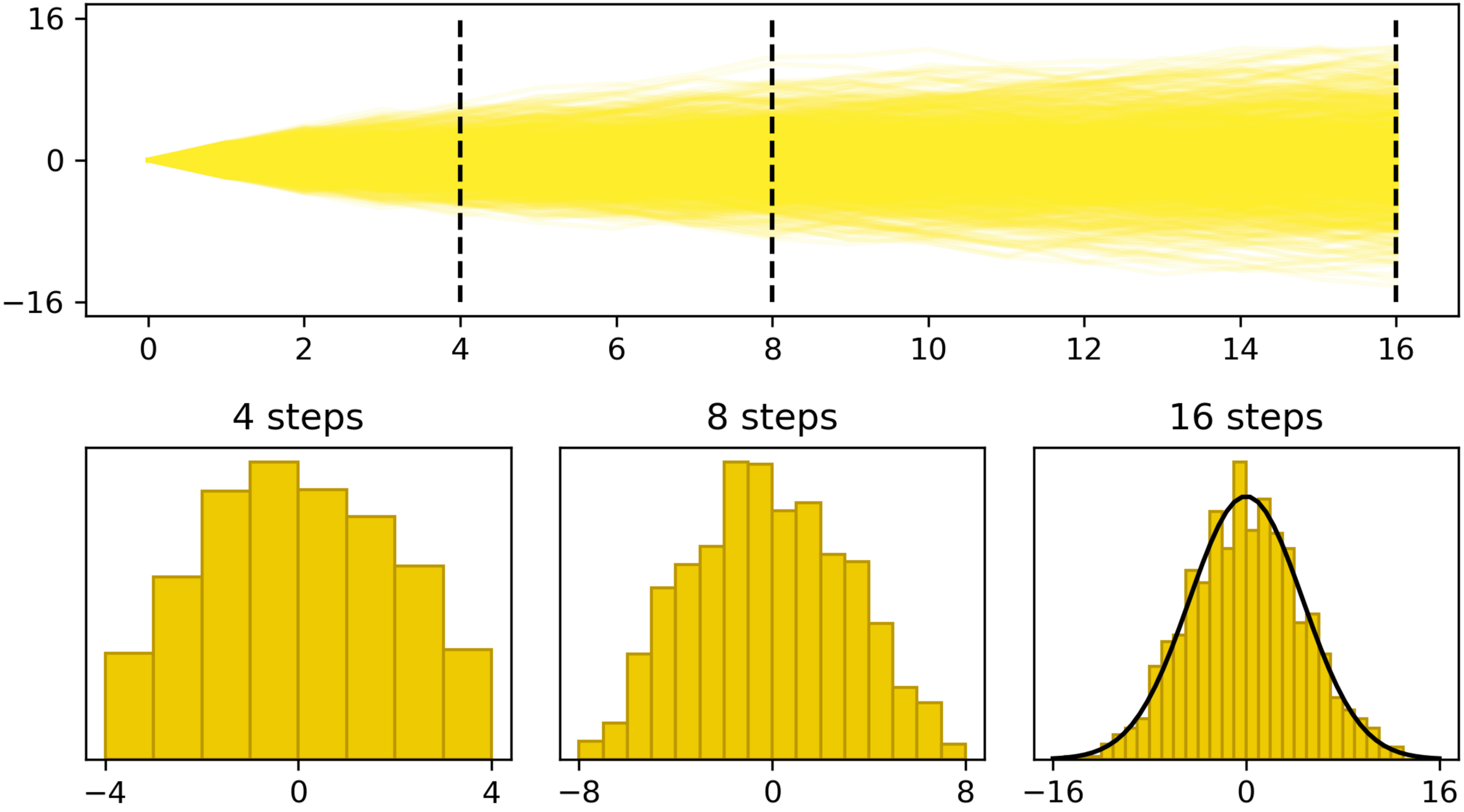
And what to do with them

- 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)

Regression with design matrix

Or what we have established so far

- Returning to our model formulation

$$\text{outcome}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta \cdot \text{predictor}_i$$

- There is no problem with adding more predictors (or functions of predictors)

$$\text{outcome}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \mathbf{x}_i \cdot \beta$$

- Which can be represented in vector matrix notation

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

$$\mu = \alpha + \mathbf{X} \cdot \beta$$

Regression with design matrix

Or what we have established so far

- Returning to our model formulation

$$\text{outcome}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta \cdot \text{predictor}_i$$

- There is no problem with adding more predictors (or functions of predictors)

$$\text{outcome}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \mathbf{x}_i \cdot \beta$$

- Which can be represented in vector matrix notation

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

$$\mu = \alpha + \mathbf{X} \cdot \beta$$

Design matrix

So what can we put into design matrix?

Complication in regression

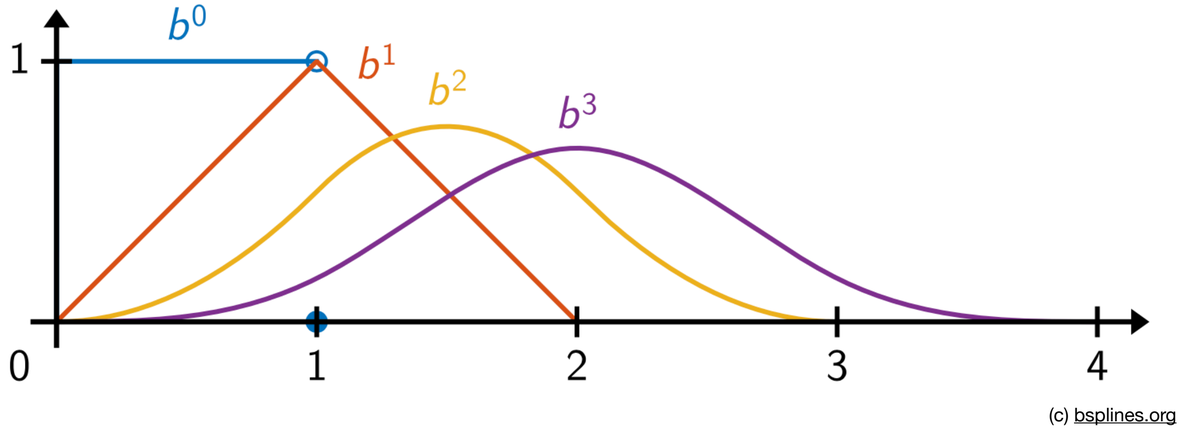
- First of all we can use multiple different predictors. We won’t talk about this in detail, but you will be given material to read.
- We have used polynomials already
- Another options are B-splines

B-spline approximation

Making function out of pieces

- Splines are functions on so called compact support
- That means that function outside certain interval (compact support) is equal to zero.
- B-splines (basis splines) are polynomials (or similar to them) on the support, while from order 1+ they are zero at the edges.

Cardinal B-splines



(c) bsplines.org

B-splines are defined by **knots**, that is point spanning on the support.

B-spline approximation

So what with knots?

- We locate knots in the are of interest and select the orders polynomial consisting of B-splines.
- Full basis of for example third order splines on considered interval consists of the functions that are nonzero at 4 knots and zero outside of them. If all such knots are different this function is differentiable. On the edges we need to repeat some knots (losing differentiability).

B-splines approximation

Construction of approximation

- Function of interest is approximated by a sum

$$f(x) \approx \beta_1 B_1(x) + \beta_2 B_2(x) + \beta_3 B_3(x) + \dots$$

- So generally it is very similar to regression

B-spline approximation

How to select knots

- Up to debate, if we have evenly distributed samples (for example time series samples) uniform distribution of knots might be a good idea.
- If samples are not evenly distributed, one rule of thumb is to locate knots at quantiles of the samples. This gives us more knots where there are more datapoints.

B-spline approximation

Why to do it?

- Splines can express local behavior better (localized effects might be expressed by few base functions)
- Splines result in sparse design matrices which can be exploited for large problems
- Splines can be considered a special case of Gaussian Processes.

Construction of design matrix

Go to the solution

- Situation is very similar to normal regression

$$\mu_i = \beta_1 B_1(x_i) + \beta_2 B_2(x_i) + \dots$$

- So that gives a design matrix:

$$\mathbf{X} = \begin{bmatrix} B_1(x_1) & B_2(x_1) & \cdots & \cdots \\ B_1(x_2) & B_2(x_2) & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_N) & B_2(x_N) & \cdots & \cdots \end{bmatrix}$$

Example

Cherry blossoming day

- Cherry trees blossom all over Japan in the spring each year, and the tradition of flower viewing (Hanami 花見) follows. The timing of the blossoms can vary a lot by year and century.
- We will consider dataset of 1000 years.

Photo by Bagus Pangestu