Bayesian Data Analysis

2020

Week 5: Linear regression

Jarno.Vanhatalo@helsinki.fi

The previous lecture

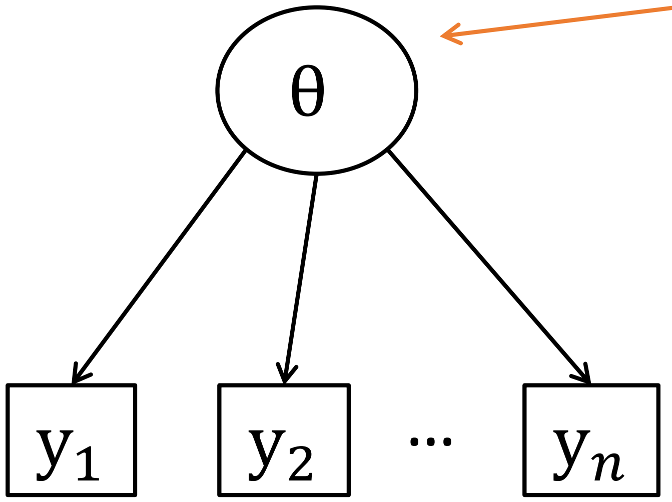
- Hierarchical models
- DAGs

Aims of the week

- Monte Carlo error
- linear regression model

Revision on exchangeability and population parameters

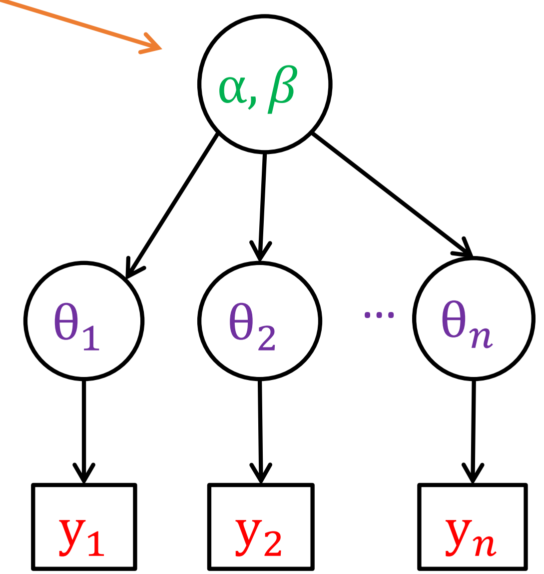
i.i.d. observations from one population



Hyperparameters of populations

Population parameters

i.i.d. observations from several ”related” populations



Estimating population parameters

- One traditional objective in statistical inference is to make generalizable statements about a *population* based on a *subset* from that population. For example,
  - Inference on the proportion of females in a population of Finns based on a random subset of Finns
  - Inference on the Proportion of females in a population of 80 year old Finns based on a random subset of 80 year old Finns
  - Inference on number of rats in Helsinki based on a number of rats in random subareas of Helsinki
  - …
- The total *population* of interest can be e.g.
  - population of all Finns
  - population of all 80 year old Finns
  - population of all rats in Helsinki
  - …
- The subset of this population is called a *sample*

Estimating population parameters

- In order to generalize the results from the sample to population the sample has to be representatitive for the population.
  - The simplest (but not necessarily the most efficient) way to achieve this is random sampling. That is, each individual in the population is equally likely to be included into the sample
- If the sample is representative to population we can assume that the sample data is coverned by *population parameter(s)* that describe the properties of the population. For example
  - If $\theta = \frac{\#\text{females}}{\#\text{females}+\#\text{males}}$ in the population of Finns, then $y \sim \text{Bin}(\theta, n)$ where y is the $\#\text{females}$ in the sample and n is the sample size

  $\rightarrow \theta$ is a population parameter
  - If μ represents the mean height of Finns and σ^2 the variance of individual heights around that mean, then $y_i \sim N(\mu, \sigma^2), i = 1, \dots, n$ where y_i is the height of i'th individual in a sample of size n

  $\rightarrow \mu$ and σ^2 are population parameters

Estimating population parameters

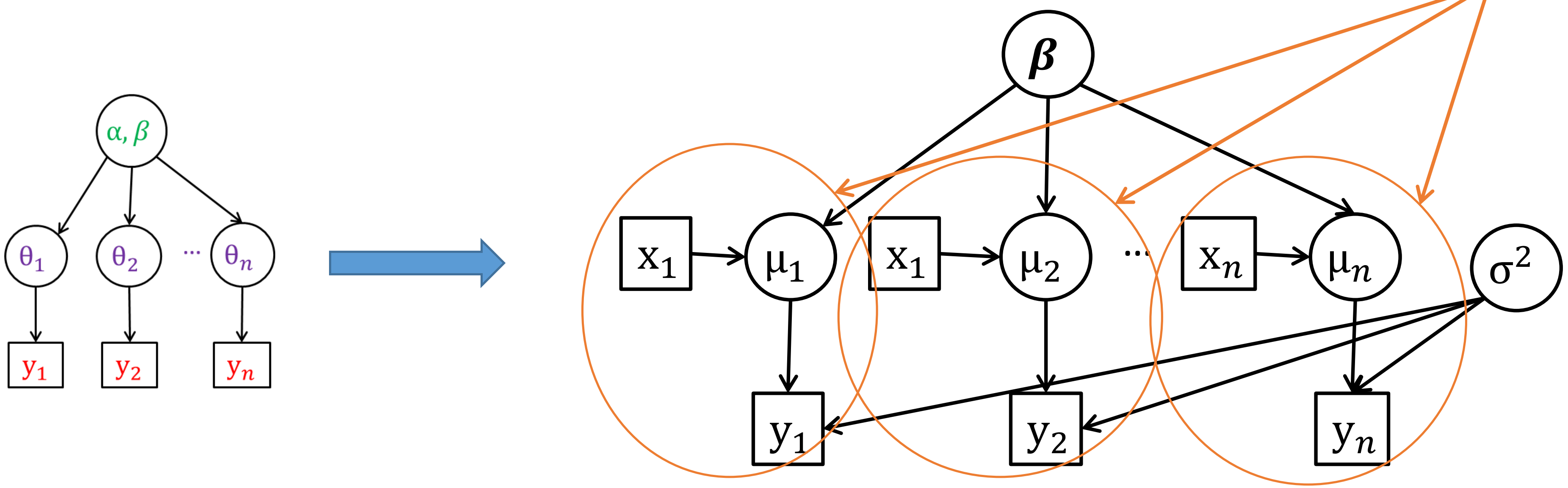
- Even though population parameter inference seems rather elementary statistical application understanding it is fundamental for more advanced models such as regression and hierarhical models.
- Estimating population parameters is also directly linked to assumptions related to exchangeable variables (week 4) and prediction.
- An example on prediction:
  - Consider we want to estimate how many Finns would want to buy a brand new cell phone. We assume their willingness to buy that cell phone depends on their age and whether they are within 5G network.
  - We make a pilot study by sampling people from different age groups and regions of Finland (either with or without 5G). This pilot study provides us a sample from the population of all Finns.
  - Let's denote by $\theta_{a,G}$ the fraction of those who would buy the phone among all Finns of age a that either have ($G = 1$) or do not have ($G = 0$) 5G in their region. Our pilot study now provides information on these population parameters $\theta_{a,G}$.
  - The prediction for the total Finnish population is then dependent on the posterior distribution of $\theta_{a,G}$ together with the information of number of Finns in each of the age-4G groups.
  - If we had not sampled all the age-4G groups we could use, for example, regression model to predict $\theta_{a,G}$ at those unsampled groups

Introduction to regression models

- ”Many studies concern relation among two or more observables. A common question is: how does one quantity, y, vary as a function of another quantity or vector of quantities, x?”
- Quantity y is response or outcome variable.
- Variables $x=(x_1,\ldots,x_k)$ are explanatory variables.
- Usual context: set of units $i=1,\ldots,n$ on which y_i and $x_{i1},\ldots,x_{ik}$ are measured.

Exchangeability and covariates

- If covariates x_i can be attached to observations y_i we can build exchangeable model for pairs (x_i, y_i)



For example, the length of a Finnish person with covariates of:
- Age
- Sex

Note! Covariates are observations as well.

Conditionally independently distributed

- Observations $y = \{\mathrm{y}_1, \dots \mathrm{y}_n\}$ from

$$\mathrm{y}_i | \mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$$
$$\mu_i = c + a x_{i,1} + b x_{i,2}$$

- Joint distribution

$$p(y|\mu, \sigma^2) = N(\mathrm{y}_1|\mu_1, \sigma^2) \times \dots \times N(\mathrm{y}_n|\mu_\mathrm{n}, \sigma^2)$$
$$= \prod_{i=1}^{n} N(\mathrm{y}_i|\mu_i, \sigma^2)$$

Non-Identical but independent distributions given $\boldsymbol{x}$

Normal linear model

- Observation model $y_i|\mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$
- The mean is thought to depend on measured variables, e.g.
$$\mu_i = c + ax_{i,1} + bx_{i,2}$$
- In matrix notation
$$\mu_i = \boldsymbol{\beta}'\boldsymbol{x}_i$$
  - $\boldsymbol{\beta}' = [c, a, b]$
  - $\boldsymbol{x_i} = \begin{bmatrix} 1 \\ x_{i,1} \\ x_{i,2} \end{bmatrix}$
- The aim is to infer the parameters of the model
- Do transformations for the data if needed (assumption about linear relationship seems somehow reasonable)

On covariates

- Covariates
  - $\mu_i = c + ax_{i,1} + bx_{i,2} = \boldsymbol{\beta}'\boldsymbol{x}_i$
- It is usually good practice to ”normalize” the covariates

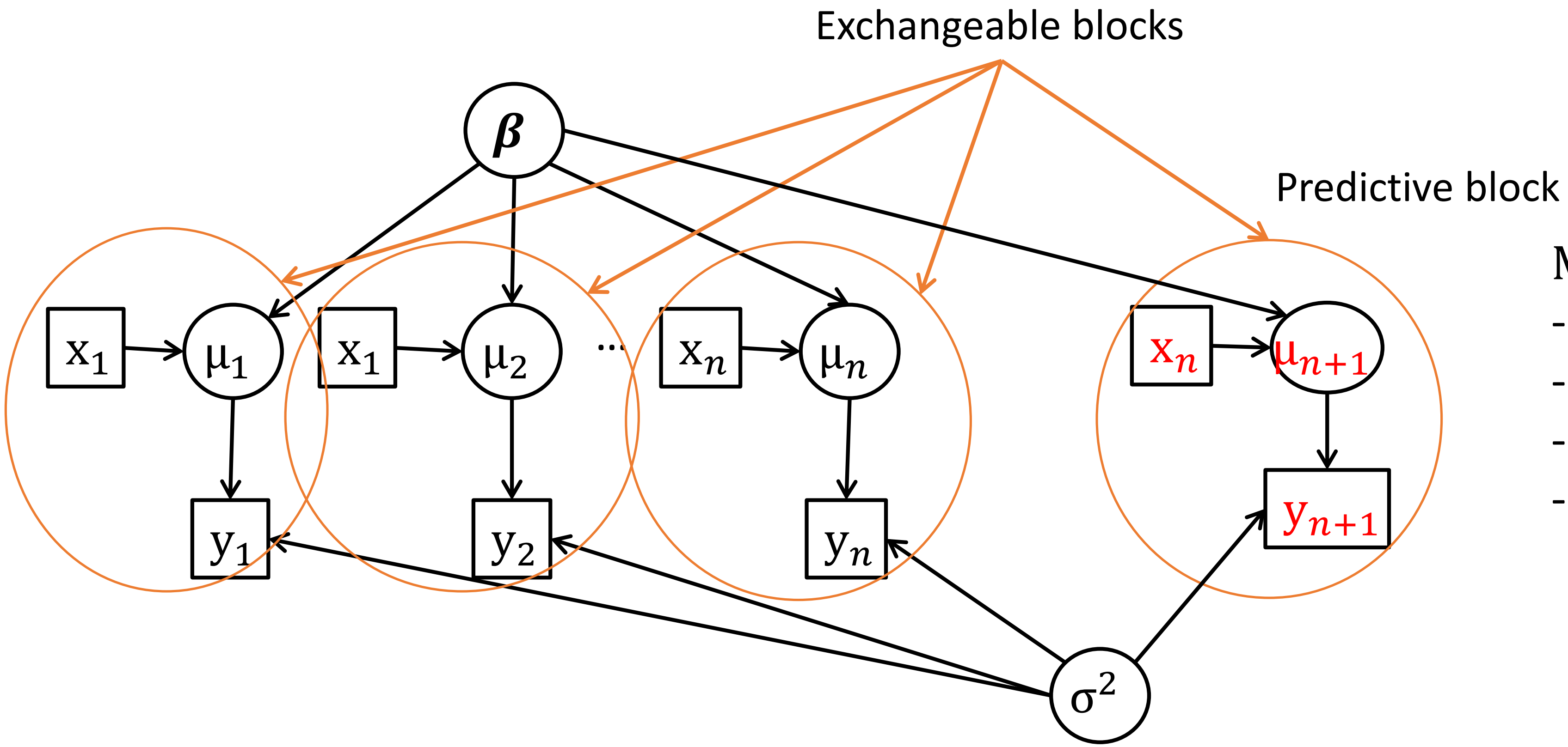
$$x_{i,1} \longrightarrow (x_{i,1} - \text{mean}\,(x_{\cdot,1}))/\text{std}(x_{\cdot,1})$$

  - Helps interpretation of the weights
  - Keeps numerics stable

Standardizing Predictors and Outputs

- https://mc-stan.org/docs/2_20/stan-users-guide/standardizing-predictors-and-outputs.html

Prediction with linear regression



Monte Carlo approximation

- Sample $\beta^s, \sigma^s \sim p(\beta, \sigma \mid \boldsymbol{y}, \boldsymbol{x})$
- calculate $\mu_{n+1}^s = \beta^s x_{n+1}$
- Sample $\tilde{y}_{n+1}^s \sim p(\tilde{y}_{n+1} \mid \mu_{n+1}^s, \sigma^s)$
- Repeat

Monte Carlo error

- Monte Carlo method is approximative
  - It introduces approximation error (uncertainty contributed by having only a finete number of simulation draws)
- Monte Carlo error in posterior mean
  - If S is large we can assume that $\frac{1}{S}\sum_{i=1}^{S} \theta^i$ is Gaussian distributed with
    - Mean $\mathrm{E}[\theta|y]$
    - variance $\mathrm{Var}[\theta|y]/S$
    - (central limit theorem)
  - If S=100 the variance of estimate $E[\theta|y]$ is 1% of the posterior variance of θ
    - Thus, Monte Carlo error is negligible for the posterior mean

Monte Carlo error

- Posterior probability

$$p(a < \theta < b|y) \approx \frac{1}{S}\sum_{i=1}^{S} I\left(a < \theta^i < b\right)$$

  - $I\left(a < \theta^i < b\right)$ are binomially distributed with probability $p\left(a < \theta^i < b|y\right)$
  - $\mathrm{Var}[I(.)] = p(1-p)$
  - Variance of the Monte Carlo estimate is $p(1-p)/S$
- Let's investigate the accuracy with coefficient of variation

  CV = Std[ ] / E[ ], where $\mathrm{Std} = \sqrt{\mathrm{Var}}$

- If S=100 and $p \approx 0.5$

  CV = $\frac{\sqrt{\frac{p(1-p)}{S}}}{p} = \frac{\sqrt{\frac{0.5(1-0.5)}{100}}}{0.5} = 10\%$
- We would need 10000 samples for 0.005/0.5 = 1% accuracy
  - Approximating small probabilities would require a lot more samples

Monte Carlo error

- Monte Carlo approximates the real quantity. E.g.
  - mean
  - Probability
- Monte Carlo estimates vary
  - > uncertainty about the real quantity
- Compare to measuring a weight in laboratory
  - take an inaccurate scale -> lot of variation between weightings
  - Take an accurate scale -> little variation between weightings
  - > Lot of Monte Carlo samples (large S) corresponds to accurate scale

Monte Carlo error in MCMC

- since the samples correlate the chain contains less information than same amount of independent samples
- Effective number of samples
  - (roughly) the number of uncorrelated samples needed to produce the same accuracy with the correlated samples
  - "Effective number of samples" ≤ "number of samples"
- See BDA pages 286-288

This week

- Gaussian model
- Regression model – linear model

Next week

- Generalized linear model
- Monte Carlo error