Applied Bayesian Modeling
Module 1: Introduction

Leontine Alkema, lalkema@umass.edu

Fall 2022

Applied Bayesian Modeling
Module 1: Introduction

Leontine Alkema, lalkema@umass.edu
Fall 2022

Motivating example

- Radon is a naturally occurring radioactive gas. Its decay products are also radioactive; in high concentrations, they can cause lung cancer (several 1000 deaths/year in the USA).
- Radon levels vary greatly across US homes.
- Data:
  - Radon measurements in over 80K houses throughout the US (we focus on Minnesota)
  - Possible predictors: floor (basement or 1st floor) in the house, soil uranium level at county level.
- Ultimate goal: predict radon levels for a non-sampled house in Minnesota (using a Bayesian hierarchical regression model).

Some analysis of the radon data

- Suppose you fit a regression model

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim N(\mu_i, \sigma^2), \\
\mu_i &= \beta_0 + \beta_1 x_i^{(\text{floor})} + \beta_2 x_i^{(\text{ura\_county})},
\end{aligned}
$$

  - $y =$ log(radon measurements);
  - $x_i^{(\text{floor})} = 1$ for 1st floor measurements, 0 basement;
  - $x_i^{(\text{ura\_county})} =$ log-uranimum in the county (centered).
- Results from a frequentist/traditional analysis:

term	estimate	std.error	conf.low	conf.high
(Intercept)	1.34	0.03	1.28	1.39
floor	-0.64	0.07	-0.78	-0.51
ura_county	0.79	0.07	0.65	0.93

- Finding for β_0, mean log-radon in basement in county with average log-uranium levels:
  point estimate is 1.34, 95% confidence interval is (1.28, 1.39).

Interpretation of confidence intervals

- Finding for β_0, mean log-radon in basement in county with average log-uranium levels:
  point estimate is 1.34, 95% confidence interval is (1.28, 1.39).
- Can you state: There is a 95% probability that β_0 is between 1.28 and 1.39.
- Nope (or "**NO NO NO ABSOLUTELY NOT!!!!**")
  - Instead: If repeated samples were taken and the 95% confidence interval was computed for each sample, 95% of the intervals would contain β_0.
- Would you **like** to construct an interval for which you can make that statement?
  Congrats, you already think like a Bayesian!

Thinking Bayesian-ly: Bayesian usage of probability

"Probabilistic reasoning ...
merely stems from our being uncertain about something.

It makes no difference whether the uncertainty relates to an unforeseeable future, or to an unnoticed past, or to a past doubtfully reported...

The only relevant thing is uncertainty - the extent of our own knowledge and ignorance.

The actual fact of whether or not the events considered are in some sense determined, or known by other people... is of no consequence."

- de Finetti (1974).

Bayesian use of probability statements

- Bayesians use probability statements to reflect a state of knowledge, i.e.
  A about an outcome in an unpredictable experiment, i.e.
    - Prob(coin toss comes up head)
    - Prob(person sampled from population is greater than 5 feet)

  B to summarize the likelihood of a statement being true, i.e.,
    - Prob(there is life on Mars)

  C some unobservable quanity (parameter) that is used to in a model to describe the data
    - Regression coefficients, Prob(β in some interval)
    - Population mean height θ, Prob($\theta > 5$ feet)
- Traditional/frequentist usage of probability is limited to describing the relative frequency of an outcome in an infinitely repeatable but unpredictable experiment.
  Question: Does that include A, B, and/or C?
  Answer: only A, not B or C

Bayesian versus frequentist thinking

- Frequentist
  - Parameter is a fixed but unknown quantity
  - Probability: to describe the relative frequency of an outcome in an infinitely repeatable but unpredictable experiment
  - Uncertainties: typically about the distribution of the data, holding the parameter fixed
- Bayesian:
  - Parameter is a random variable
  - Probability statements reflect a state of knowledge
  - Uncertainties: typically about the distribution of the parameter, holding the data fixed

Bayesian inference

- Goal: draw statistical conclusions about parameters of interest using probability statements.
- General approach for some outcome of interest θ, i.e. regression coefficient, is based on learning via Bayes' rule
  - start off with *prior* probability distribution to quantify information related to θ
  - collect data, and use Bayes' rule to update the prior into the posterior distribution
  - use posterior distribution to draw conclusions

Motivation example (ctd)

- Frequentist inference about β_0 in the radon example:
  point estimate is 1.34, 95% confidence interval is (1.28, 1.39)
- Bayesian inference about β_0 in the radon example (details to follow):
  point estimate 1.34, 95% credible interval (1.28, 1.39)
- Here we get the same (data-driven) interval but with different interpretations: based on the 95% credible interval, we conclude that there is a 95% probability that β_0 is between 1.28 and 1.39.

Bayesian models

- Bayesian modeling or Bayesian data analysis refers to doing statistical analysis of data “Bayesian-ly”;
  - to use probability distributions for quantifying uncertainty in data and model parameters, and
  - to use Bayes’ rule for drawing inference
- Why go Bayesian?
  - Uncertainty assessments automatically incorporate uncertainty about the true values of all parameters
  - (personal opinion/experience) Once you are familiar with the basic ideas in Bayesian inference, and can work with software such as Stan, fitting more complex models is much easier in the Bayesian framework as opposed to the frequentist approach.
- This course on Applied Bayesian modeling:
  - how to construct Bayesian models to relate data to scientific questions,
  - to fit such models fitting using statistical programs (R and STAN),
  - to interpret model results,
  - to check model assumptions (follow Bayesian workflow to model building).

Applied Bayesian Modeling: Core modules

- Introduction to Bayesian modeling
- Bayesian inference
  - 1 discrete outcome
  - 1 continuous outcome (everything's normal)
  - Deeper dive into the normal set-up; bias-variance trade-off
- Computation, model fitting
  - Sampling-based approach to inference when posterior is not available in closed form
  - How to sample from a posterior density
  - Let's stan: what to check and what to consider tuning?
- Hierarchical models I and II
- Expanding our model universe
- Model checking and validation I and II; Bayesian work flow

Course organization

- Focus on learning through doing!
- Flipped classroom
  - Watch short lecture(s) prior to class time
  - Spend class time on exercises/HW
  - Undertake a substantial project in Bayesian modeling
- Main material is organized into core modules with specific learning objectives:
  - Core modules are meant to be a minimally sufficient and time efficient introduction to applied Bayesian modeling, focused on the specific learning objectives only
  - See course website for additional readings/resources

Additional details for Fall 2022 BIOSTATS 730

- Please check syllabus (and schedule) for details, and use our slack workspace for discussion/questions/etc
- Grading: HWs, Exams, Project, Participation
- HWs:
  - We typically work on a HW on Tu and Th in class, it is due the following Monday (with no-cost-but-not-advised extension to Wed).
  - Collaboration on HWs is allowed but you need to write your own solutions.
- Proposed exam dates:
  - Closed book: Tu 10/11, 230-430PM (note later end compared to 345pm class time)
  - Take home: 24 hour period out of Monday 11/14 - Wed 11/16 (no class that Tu)
- Project presentations, during class time:
  - Tu 11/18
  - Tu 12/6 and Th 12/8 (last week of class)
- Participation: watch videos prior to class, engage during class and asynchronous on slack