Bayesian Statistics

"Hey, you know?
Oh, I don't know!
I know but I don't know"
-Blondie

Collin Cademartori

April 17, 2020

Bayes Rule

- Let A and B be two events.
- Recall the probability of A given B is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}$$

- "The proportion of the times that B happens that A also happens."
- Bayes rule allows us to compute one conditional probability in terms of the opposite

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Regression

- Recall the basic linear regression model:

$$y_i = \alpha + \beta_1 x_i^1 + \cdots + \beta_k x_i^k + \epsilon_i$$

- The i subscripts index our observations of these variables
- y is some variable we want to predict
- $x^1, \ldots, x^k$ are variables we believe predict y
- ϵ captures variance in y not predicted by the xs
- ϵ often thought of as random noise or measurement error
- We assume the ϵ_i are independent across observations i
- Also assume ϵ_i have the same distribution

Regression (cont.)

- Linear regression is a special case of models of form

$$y_i = f\left(x_i^1, \ldots, x_i^k\right) + \epsilon_i$$

- Here f may be a more complex (i.e. non-linear) function
- The ϵ_i may not be independent
  - Could be correlated in time
  - Or correlated within groups
- The ϵ_i may have different distributions
  - Variance could vary with x
  - Or could vary by group
- Linear regression can be extended to accomodate these complexities

What is Bayesian Statistics?

- In classical statistics, we want to estimate parameters
- Think of parameters as fixed but unknown
- Then observe data generated from parameters
- Use the data to construct an estimator
- Estimators justified by models for data given parameters
- Regression example: $y_i = \alpha + \beta x_i + \epsilon_i$
  - Assume errors ϵ_i have normal distribution
  - Then we can derive the maximum likelihood estimator
  - These estimates make data as probable as possible

What is Bayesian Statistics?

- How is Bayesian statistics different?
- We think of the unknown parameters θ as random
- In a Bayesian model we have a "sampling distribution":

$$x \sim p(x \mid \theta)$$

  Describes how observed data are generated given parameters
- And we have a "prior distribution" :

$$\theta \sim p(\theta)$$

  Describes process that generated unknown parameters

The Bayesian Mindset

- Why introduce added complexity into the model?
- Bad news: need to know something about θ
- Good news:
  - Can include information we have about θ
  - Prior assumptions often interpretable (thus checkable)
  - Natural way to add complexity to models
- How to minimize the bad news?
  - Can choose weak prior if we have little information
  - Simulations from model can reveal bad assumptions

The Punchline

- But how do we use our model to estimate θ?
- In maximum likelihood, wanted θ that could explain x
- Now that θ is random, we can make sense of "θ given x"
- Bayes rule tells us how to compute this conditional probability:

$$p(\theta \mid x) = \frac{\overbrace{p(x \mid \theta)}^{\text{sampling dist.}} \times \overbrace{p(\theta)}^{\text{prior dist.}}}{p(x)}$$

- Notice that $p(x)$ does not depend on θ
- Think of x as fixed observed data, so $p(x)$ constant

The Distribution $p(x)$

$$p(\theta \mid x) = \frac{\overbrace{p(x \mid \theta)}^{\text{sampling dist.}} \times \overbrace{p(\theta)}^{\text{prior dist.}}}{p(x)}$$

- Because $p(\theta \mid x)$ is a probability distribution:

$$1 = \int_\theta p(\theta \mid x) d\theta = \frac{1}{p(x)} \int_\theta p(x \mid \theta) p(\theta) d\theta$$

- Rearranging, we getting

$$p(x) = \int_\theta p(x \mid \theta) p(\theta) d\theta$$

- So $p(x)$ is completely determined by $p(x \mid \theta)$ and $p(\theta)$

Posterior Distributions

- The upshot:
  - Specify sampling distribution $p(x \mid \theta)$
  - Specify prior distribution $p(\theta)$
  - These uniquely determine a posterior distribution $p(\theta \mid x)$
  - Up to a constant, this is just $p(x \mid \theta)p(\theta)$
- Posterior distribution $p(\theta \mid x)$ relates data to unknown parameters θ
- Posterior eliminates need to design estimators separately from the model
- Also quantifies our uncertainty for free
  - Very wide distribution $\rightarrow$ data don't tell us much about θ
  - Very narrow distribution $\rightarrow$ data precisely identify θ
  - Other possibilities: What if posterior multi-modal?

Priors on Priors

- How do we choose $p(\theta)$ in practice?
- Often based on knowledge of process generating θ
- If we have uncertainty about that process, often natural to introduce more parameters λ
- So we have prior $\theta \sim p(\theta \mid \lambda)$
- Then need a hyper-prior for $\lambda \sim p(\lambda)$
- If we don't know much about λ, can just use weak prior
- This leads to a general straegy:
  - Add higher-level parameters to model how lower-level parameters are generated
  - Continue until we don't know how highest-level parameters generated
  - Place weak prior on highest-level parameters

Hierarchical Models

- Models constructed this way are called hierarchical models
- Can think of modeling assumptions sequentially
- Generate highest level parameters λ $p(\lambda)$ first
- Given these, generate $p\theta \sim (\theta \mid \lambda)$ next
- Continue until we generate data x
- Can be shown that higher level parameters have less influence on data
- This justifies weak priors on highest level parameters λ

Example: Eight Schools

- Let's look at a famous toy problem
- Eight schools conduct experiments to assess their SAT coaching programs
- They randomly assign students to receive or not receive coaching
- They then estimate the effect of the coaching programs on SAT scores
- The resulting data are estimated average effects and standard errors
- The average effect is the average difference between a student in the treatment and control group

Eight Schools Continued

The data from the 8 schools looked like the following:

Avg. Effect	Std. Err.
28	15
8	10
-3	16
7	11
-1	9
1	11
18	10
12	18

Can we conclude whether any of these coaching programs had an effect on SAT scores?

A Model for the 8 Schools: Sampling Distribution

- First, for $i = 1, \ldots, 8$, let x_i be the average treatment effect and s_i the standard error for school i
- Sampling distribution:

$$p(x_i \mid \theta_i) = \text{normal}\left(\theta_i, s_i\right)$$

- θ_i is the true treatment effect for school i.
- x_i could differ from θ_i because:
  - SAT test taken imperfect representative of average SAT test
  - Student performance on that day imperfect measure of long-run performance
- Hence the s_i captures the measurement error of the particular experiment schools carried out

A Model for the 8 Schools: Prior Distribution on θ_i

- We have introduced average treatment effect parameters θ_i
- So we need a prior distribution on them
- First question - are they independent?
- If seven of eight school show no evidence of an effect, what about the eighth?
- Seems like strong evidence that SAT coaching programs don't work in general
- So we should allow school treatment effects to be correlated
- Hierarchical modeling gives us a natural way to do this!
- Take $p(\theta_i \mid \mu, \tau) = \text{normal}(\mu, \tau)$
- So treatment effects come from a common super-population

A Model for the 8 Schools: Prior on θ_i

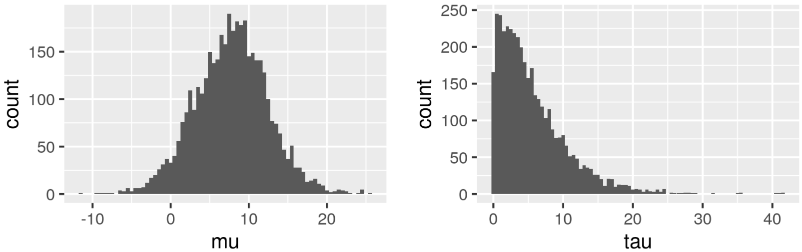
- Note that θ_i are independent *given* μ and τ
- So correlation is expressed as uncertainty about the population from which they are drawn
- If SAT coaching doesn't work, we expect $\mu \approx 0$ and τ small
- If some work and some don't, we expect τ to be larger
- So τ controls the amount of information θ_i share
  - $\tau = 0$ corresponds to all coaching effects being equal
  - $\tau \to \infty$ corresponds to completely independent effects
  - Since τ estimated from data, we allow data to inform how correlated schools should be

A Model for the 8 Schools: Hyper-Priors

- How to interpret μ and τ?
    - μ is the average of the average treatment effects θ_i
    - τ is the variance of these average treatment effects
- Note that θ_i averages over the students in a school and SAT tests they could take
- ...while μ averages over the different schools
- Don't have much information about μ and τ
- For μ, we might be agnostic about the effect of coaching
- Since score range is 1200, might think max effect no larger than 100.
- Can then choose $p(\mu) = \text{normal}\,(0, 35)$
- For τ, choose weak $p(\tau) = \text{Exponential}(30)$ distribution

8 Schools: The Posterior

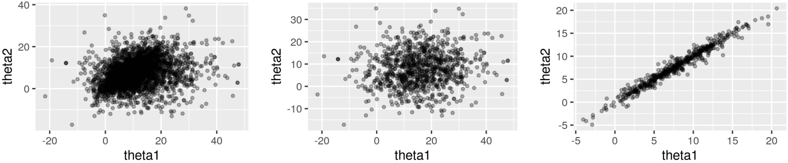
Posterior histograms for μ and τ:



The data seem to support a small positive average effect along with a bit of variation between schools. However, the data also could have been generated by the model even if the true μ value was 0. And the data are consistent with no variation between schools.

8 Schools: The Posterior

We can visualize what these hyper-parameter distributions imply for the θ_i values.



The first plot shows the posterior of (θ_1, θ_2).
The second shows (θ_1, θ_2) conditional on $\tau > 8$.
The third shows (θ_1, θ_2) conditional on $\tau < 1$.

Computing Posteriors

- So the posterior relieves us of estimator design
- But how do we actually find the posterior in practice?
- We can compute the posterior up to a constant by

$$p(x \mid \theta)p(\theta)$$

- Often we need to compute marginal distributions:

$$p(\theta_1, \theta_2) = \int_{\{\theta_i | i \geq 3\}} p(\theta \mid x) d\theta_3 \cdots d\theta_k$$

- Or averages over the posterior distribution:

$$\mathbb{E}\left[\phi(\theta) \mid X\right] = \int_\theta \phi(\theta) f(\theta \mid x) d\theta$$

Computing Posteriors

- Computing these integrals is generally really hard
- No analytical solution in most cases
- Good news: Don't have to remember integration by parts
- Bad news: Need complicated numerical procedures to approximate integrals
- Suppose we could get samples $\theta^{(1)}, \ldots, \theta^{(S)}$ from $p(\theta \mid x)$
- If they're independent, the law of large numbers says that

$$\lim_{S \to \infty} \frac{1}{S} \sum_{i=1}^{S} \phi\left(\theta^{(i)}\right) = \mathbb{E}\left[\phi\left(\theta\right) \mid X\right]$$

- The law of large numbers still works if samples are only approximately independent

Monte Carlo Simulation

- Now we just need a scheme to generate these random samples
- Modern state-of-the-art samplers are very complex
- We will look at a simple scheme that works in one dimension
- First we assume that we can sample from a uniform distribution on $[0, 1]$
- (i.e. all real numbers between 0 and 1 are equally likely)
- Can generate these with pseudo-random number generators
- These use number theory to mimic random number generation
- Widely available on all modern computers

Inverse CDF Method

- Let $U_1, \dots, U_S$ be samples from the uniform distribution
- Suppose we want samples from 1-dimensional distribution $p(x)$
- If $X \sim p(x)$, we define the cumulative distribution function (CDF) $F(x)$ by:
$$F(x) = \mathbb{P}(X \leq x)$$
- $F(x)$ is increasing, often invertible. Let $Q(x)$ be the inverse.
- We claim $Q(U_1), \dots, Q(U_S)$ are a sample from $p(x)$
- If $Y_i = Q(U_i)$ has CDF $F(x)$, it has distribution $p(x)$

Inverse CDF Method

Proof that inverse CDF method works:

$$
\begin{aligned}
F_{Y_i}(x) &= \mathbb{P}\left(Y_i \leq x\right) && (1)\\
&= \mathbb{P}\left(Q(U_i) \leq x\right) && (2)\\
&= \mathbb{P}\left(F(Q(U_i)) \leq F(x)\right) && (3)\\
&= \mathbb{P}\left(U_i \leq F(x)\right) && (4)\\
&= F(x) && (5)
\end{aligned}
$$

- (1) follows by definition of the CDF
- (2) follows by definition of $Y_i = Q(U_i)$
- (3) follows since F is increasing (so preserves inequalities)
- (4) follows since Q is inverse of F
- (5) follows since $\mathbb{P}\left(U_i \leq a\right) = a$ for $a \in [0, 1]$ (by uniformity)

A Worked Example

- Now we will look at an example modeling problem
- Consider a randomized experiment of a cholesterol medication
- Suppose there are 200 people in the trial and 100 receive the drug
- Assume the experiment uses a placebo and is double blind
- We take two measurements of each person:
  - One cholesterol measurement just before administering the drug (or placebo)
  - One cholesterol measurement two weeks after treatment

A Worked Example

- We will consider a model of the form

$$y_{i0} = \alpha + \epsilon_{i0}$$

$$y_{i1} = \alpha + \beta \times T_i + \epsilon_{i1}$$

- y_{i0} is the pre-treatment measurement for individual i
- y_{i1} is the post-treatment measurement for individual i
- α represents the baseline average cholesterol for the group
- $T_i = 1$ if individual i received the drug, $t_i = 0$ if they received the placebo
- β represents the average effect of receiving the drug
- ϵ_{i0} and ϵ_{i1} are independent error terms