Bayesian Statistics

Dustin Leatherman

December 6, 2020

Contents

1 Intro (2020/09/10)

1.1 Motivating Example

There are two students: Student A and Student B, along with an instructor. A secretly writes down a number (1,..,10) then mentally calls heads or tails.

1. The instructor flips a coin
2. If heads, A honestly tells B if the number is even or odd.
3. If A guesses H/T correctly, A tells B if their number is even or odd. Otherwise, they lie.
4. B will guess if the number is odd or even

   Let θ be the probability that B correctly guesses even or odd.

   The class (and myself) initially agreed without much discussion that 0.5 is the obvious answer. Upon further thinking on this, the probabilities breakdown in such way:

   (2): 0.5 (3): 0.5 (4): ?

   The initial logic is that its a 50/50 chance since there are two choices but there is an X-factor here with number 4. A few questions worth asking:

   1. Does B know the rules upfront? As in, are they

aware that A may or may not lie?

1. Does B see the result of the coin flip?
2. Is this done virtually or in person?

If the answer is no for 1 and 2, then 0.5 is a logical choice because they'd be guessing without much foreknowledge.

If the answer is yes for 1 and 2, then B is in on the "game" and can make a more educated guess. If A or the professor has a "tell", then that could provide information. Reading body language may also provide some information to B on the veracity of A's claim.

I would argue that θ would be > 0.5 *if* A and B know each other well enough. Which is really a great example of Bayesian vs Frequentist view points.

1.2 Frequentist Approach

Quantifies uncertainty in terms of repeating the process that generated the data many times.

1.2.1 Properties

- The parameters θ are fixed, unknown, and a constant.
- The sample (data) Y are random
- All prob. statements would be made about the randomness in the data.
-

A statistic $\hat{\theta} = Y/n$ is a statistic and is an estimator of the population proportion θ

The distribution of $\hat{\theta}$ from repeated sampling is the *sample distribution.*

1.2.2 Things a Frequentist would never say

- $P(\theta > 0) = 0.6$ because θ is not a random variable
- The distribution of θ is Normal(4.2,1.2)

- The probability that the true proportion is in the interval (0.4, 0.5) is 0.95.
- The probability that the null hypothesis is true is 0.03.

1.3 Bayesian Approach

Expresses uncertainty about θ using probability distributions. θ is still fixed and unknown.

Distribution *before* observing the data is the **prior distribution**. e.g. $P(\theta > 0.5) = 0.6$. This is subjective since people may have different priors.

Hopefully, Uncertainty about θ is reduced after observing the data.

Bayesian Interpretations differ from *Frequentist* Interpretations.

Uncertainty distribution of θ after observing the data is the **posterior distribution**.

Bayes Theorem for updating the prior

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)} \tag{1}$$

Described in words: Posterior $\propto$ Likelihood $\times$ Prior

$f(\theta|Y)$ is the posterior distribution.

Given that I have seen some data, what am I seeing now?

A key difference between Bayesian and frequentist statistics is that all inference is conditional on the single data set we observed (Y).

1.3.1 Likelihood Function

Distribution of the observed data given the parameters. This is the Same function used in a maximum likelihood analysis.

> When prior information is weak, Bayesian and Maximum Likelihood Estimates are similar.

1.3.2 Priors

Say we observed Y = 60 successes in n = 100 trials and $\theta \in [0, 1]$ is the true probability of success.

Want to select a prior that has a domain of [0, 1]

If there is no relevant prior information, we might use $\theta \sim Uni(0, 1)$. This is called an *uninformative prior*. aka a "best guess".

1. Beta

   Beta distributions are a common prior for parameters between 0 and 1.

   If $\theta \sim Beta(a, b)$, then the posterior is

   $$\theta|Y \sim Beta(Y + a, n - Y + b)$$

   $Beta(1, 1) == Uni(0, 1)$

2. Gamma Popular distribution for σ (population standard deviation)

1.3.3 Posteriors

The likelihood function $Y|\theta \sim Bin(n, \theta)$
The Uniform prior is $\theta \sim Uni(0, 1)$
The posterior is then $\theta|Y \sim Beta(Y + 1, n - Y + 1)$

1.3.4 Advantages

- Bayesian concepts (posterior probability of the null hypothesis) are arguably easier to interpret than the frequentist ideas (p-value.)
- Can incorporate scientific knowledge via the prior.
  - Even a Small amount of prior information can add stability.
- Excellent at quantifying uncertainty in complex problems.
- Provides a framework to incorporate data/information from multiple sources.

1.3.5 Disadvantages

- Less common/familiar
- Picking a prior is subjective (though there are objective priors)
- Procedures with frequentist properties are desirable.
- Computing can be slow for hard problems
- Non parametric methods are challenging

1.4 Review

Only the interesting parts are placed here. See the rest of this repo for deeper dives on other concepts.

1.4.1 Probability

Objective (associated with Frequentist)

- $P(X = x)$ is a mathematical statement
- If we repeatedly sampled X, the value that the proportion of draws equal to x converges is defined as $P(X = x)$

Subjective (associated with Bayesian)

- $P(X = x)$ represents an individual's degree of belief
- Often quantified as the amount an individual would be willing to wager that X will be x.

A Bayesian Analysis uses both of these concepts.

1.4.2 Uncertainty

Aleatoric (def: indeterminate) uncertainty (likelihood)

- Uncontrollable randomness in the experiment

Epimestic (def: involving knowledge) uncertainty (prior/posterior)

- Uncertainty about a quantity that could be theoretically

A Bayesian Analysis uses both of these concepts

1.4.3 Probability vs Statistics

The common sense, I like the way this is phrased.

Probability is the forward problems

- We assume we know how the data are being generated and computer the probability of events.

  For example, what is the probability of flipping 5 straight heads if the coins are fair?

Statistics is the inverse problem

- We use data to learn about the data-generating mechanism

  For example, if we flipped five straight heads, can we conclude the coin is biased?

2 Probability & Introduction to Bayes (2020/09/17)

if x and y are independent, then the following is true

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

Cannot use $f(x,y)$ as PMF because $\sum_1^Y f(x,y) = f(x) \neq 1$. Need to scale by marginal probability in order to sum to 1 and thus be a proper PMF/PDF.

	1	2	3	4	5	Total [p(y)]
US	.0972	.0903	.0694	.0069	.0069	.2708
Not US	.3194	.1319	.1389	.1181	.0208	.7292
Total [p(x)]	.4167	.2222	.2083	.1250	.0278	1

show that x and y are dependent
$P(x = 1) = 0.4167$
$P(y = 1) = 0.2708$
$P(x = 1) \times P(y = 1) = 0.4167(0.2708) = 0.1128$
$P(x = 1, y = 1) = 0.0972 \neq 0.1128$ so dependent!

2.1 Calculating the Posterior Analytically

2.1.1 Using an Arbitrary PDF

1. Find Joint Probability (f(x,y))

$$
\begin{aligned}
P(x>7, y>40) &= \int_7^{10}\int_{40}^{50} 0.26exp(-|x-7|-|y-40|)\ dx\ dy \\
&=0.26\ \int_7^{10}\int_{40}^{50} exp(-x+7-y+40)\ dx\ dy \text{ (Since only interested in positive values)} \\
&=0.26\ \int_7^{10}\int_{40}^{50} exp(-(x-7)\ exp(-(y-40))\ dx\ dy \\
&=0.26\ \int_7^{10}\int_{0}^{10} exp(-(x-7)\ exp(-u)\ dx\ du \\
&=0.26\ \int_7^{10}\int_{0}^{10} exp(-(x-7)\ [-exp(-u)]_0^{10}\ dx\ du \\
&=0.26(1-e^{-10})\ \int_7^{10} exp(-(x-7)dx \\
&=0.26(1-e^{-10})(1-e^{-3}) \approx 0.247
\end{aligned} \tag{2}
$$

1. Find Marginal Probability over the Data $f_X(x)$

$$
\begin{aligned}
f_X(x) &= \int_{20}^{50} 0.26 + e^{-|x-7|-|y-40|}dy \\
&=0.26e^{-|x-7|}\int_{20}^{50} e^{-|y-40|}dy \\
&=0.26e^{-|x-7|}[\int_{20}^{40} e^{-(40-y)}dy + \int_{40}^{50} e^{-(y-40)}dy] \\
&=0.26e^{-|x-7|}[\int_{20}^{0} -e^{-u}du + \int_{0}^{10} e^{-u}du] \\
&=0.26e^{-|x-7|}[1-e^{-20}+1-e^{-10} \approx 2] \\
&=0.52e^{-|x-7|}\ \forall\ x \leq x \leq 10
\end{aligned} \tag{3}
$$

1. Calculate Conditional Probability

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{2}e^{-|y-40|}$$

If integrating over an absolute value, break up the integral into two integrals: the first over the negative domain of the integration, the second over the positive domain.

2.1.2 Using Normal Distribution

1. Find Marginal Probability

$$
\begin{aligned}
f(x) =& \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\ exp(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)})\ dy \\
=& \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-x^2/2(1-\rho^2)}\ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} exp(-\frac{y^2-2\rho xy}{2(1-\rho^2)})\ dy \text{ (Move x's out of integral. Arrange term} \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-x^2/2(1-\rho^2)}\ \int_{-\infty}^{\infty} \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi(1-\rho^2)}} exp(-\frac{y^2-2\rho xy+\rho^2x^2-(\rho x)^2}{2(1-\rho^2)})\ dy \\
=& \frac{1\ \sqrt{1-\rho^2}}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-x^2/2(1-\rho^2)}\ e^{\frac{\rho x^2}{2(1-\rho^2)}}\ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} exp(-\frac{(y-\rho x)^2}{2(1-\rho^2)})\ dy,\ N(\rho x, 1-\rho^2) \\
=& \frac{1}{\sqrt{2\pi}} e^{-0.5\ \frac{x^2-\rho^2x^2}{1-\rho^2}} \\
=& \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}, X \sim N(0,1)
\end{aligned}
\tag{4}
$$

1. Assume Joint Normal PDF
2. Find Conditional probability

$$
\begin{aligned}
f(y|x) =& \frac{f(x,y)}{f_X(x)} \\
=& \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} exp(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)})}{\frac{1}{\sqrt{2\pi}}\ exp(-\frac{x^2}{2})} \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)} + \frac{x^2}{2}) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{x^2+y^2-2\rho xy}{1-\rho^2} - x^2]) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{x^2+y^2-2\rho xy-(1-\rho^2)x^2}{1-\rho^2}]) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{y^2-2\rho xy-\rho^2x^2}{1-\rho^2}]) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{(y-\rho x)^2}{1-\rho^2}]),\ y|x \sim N(\rho x, 1-\rho^2)
\end{aligned}
\tag{5}
$$

2.2 Bayes Theorem

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

How do you know you are using Bayes Rule?
Given $P(y|\theta)$, want to find $P(\theta|y)$

- Bayesians quantify uncertainty about fixed but unknown parameters by treating

them as random variables.

- This requires that we set a prior distribution $\pi(\theta)$ to summarize uncertainty before observing the data.
- The distribution of the observed data given the model parameters is the *likelihood function*, $f(Y|\theta)$
  - The likelihood function is the most important piece of a Bayesian Analysis because it links the data and the parameters.

2.3 Bayesian Learning

The posterior distribution $P(\theta|Y)$ summarizes uncertainty about the parameters given the prior and data.

Reduction in uncertainty from prior to posterior represents **Bayesian Learning**

Bayes Theorem (again):

$$P(\theta|Y) = \frac{f(Y|\theta)\pi(\theta)}{m(Y)}$$

$m(Y) = \int F(Y|\theta)\pi(\theta)d\theta$: marginal distribution of the data and can usually be ignored.

2.4 Subjectivity

Choosing Likelihood function and a prior distribution are subjective.

If readers disagree with assumptions, findings will be rejected so assumptions must be justified theoretically and empirically.

3 Summarizing a Posterior Distribution (2020/09/24)

3.1 SIR Model

Susceptible-Infected-Recovered

At time t , $S_t + I_t + R_t = N$ where N is the population.

States evolved according to the following differential equations

$$\begin{aligned} \frac{dS_t}{dt} &= -\beta\frac{S - tI_t}{N} \\ \frac{dI_t}{dt} &= \beta\frac{S_t I_t}{N} - \Gamma I_t \end{aligned} \tag{6}$$

β: Controls rate of new infections

Γ: Controls recovery rate

We will use a discrete approx to these curves with hourly time steps.

So? $dt = \frac{1}{24}$

Goal: Fit SIR Model for given values of β and Γ

3.2 Summarize a univariate Posterior with Beta-Binomial

Posterior = Likelihood × Prior

Say there is a parameter θ

Likelihood: $Y|\theta \sim Bin(N, \theta)$

Prior: $\theta \sim Uni(0, 1) \equiv Beta(1, 1)$

Posterior: $\theta|Y \sim Beta(Y + a, N - y + b)$

> Peak of the Posterior is the MLE of the Likelihood function when using an uninformative prior.

3.3 MAP Estimator

Posterior Mode is call the max a posterioiri (MAP) estimator

$$\hat{\theta} = \underset{\theta}{argmax}\ P(\theta|y) = \underset{\theta}{argmax}\ log[f(Y|\theta)] + log[\pi(\theta)] \tag{7}$$

if prior is uniform, MAP is MLE assuming $Y|\theta \sim Bin(\theta, n)$.

3.4 Uncertainty Measures

Posterior Std. Dev. is one measure of uncertainty

- If approx Gaussian, can use empirical rule
- Analogous but fundamentally different than std error.
  - Std err is the standard deviation of $\hat{\theta}$'s sampling distribution

Do not call them call them **confidence** intervals. Called **Credible** Intervals in Bayesian Statistics.

Interval (l, u) is $100(1 - \alpha)\%$ posterior credible interval if $P(l < \theta < u|Y) = 1 - \alpha$

Interpretation: "Given the data and the prior, the probability that θ is between l and u is 0.95."

> Confidence interval interpretation:
>
> With 95% Confidence, θ is between l and u.

A Bayesian Posterior is a distribution for $\theta|Y$ whereas the sampling distribution is for $\hat{\theta}$. While their expected values both represent the true mean, the sampling distribution is not a distribution of θ hence why "Confidence" is used when in the interpretation. The Bayesian Posterior is a distribution of θ so the posterior can be used for the interpretation.

3.4.1 Credible Sets

Not unique.

Let q_τ be the τ quantile of the posterior of the posterior such that $P(\theta < q_\tau|Y) = \tau$. Then $(q_{00}, q_{0.95})$, $(q_{0.01}, q_{0.96})$, etc. are all valid 95% credible sets.

Equal-Tailed intervals: $(q_{\alpha/2}, q_{1-\frac{\alpha}{2}})$

Highest posterior density interval searches for the smallest interval that contains the proper probability

3.5 Hypothesis Tests

Conducted by computing posterior prob of each hypothesis.

$$P(\theta < 0.5|Y) = \int_0^{0.5} P(\theta|Y)d\theta$$

analogous but different than a p-value.

p-value: Assuming the null hypothesis is true, the probability we got X or a value more extreme is Y.

Bayesian Hypothesis Test: Given the prior and the data, the probability the null hypothesis is true is Y.

3.6 Monte Carlo Sampling

A useful tool for summarizing a posterior.

In MC sampling, we draw S samples from the posterior;

$$\theta', ..., \theta^{(s)} \sim P(\theta|Y)$$

and use these samples to approx the posterior.

3.6.1 Transformations

MC sampling facilitates studying the **transformations** of parameters.

For example, the odds corresponding to θ are $\gamma = \frac{\theta}{1-\theta}$

$$\gamma^{(1)} = \frac{\theta^{(1)}}{1-\theta^{(1)}}, ..., \gamma^{(S)} = \frac{\theta^{(S)}}{1-\theta^{(S)}} \tag{8}$$

How to approximate the posterior mean and variance of γ?

Transform the odds and use the draws to approximate θ's posterior!

3.7 Summarizing Multivariate Posteriors

Univariate posteriors captured by a simple plot. Not easy or impossible to do with multivariate posteriors.

Let $\theta = (\theta_1, ..., \theta_p)$.

Ideally, we reduced to the univariate marginal posteriors. Then the same ideas for univariate models apply

$$P(\theta_1|Y) = \int ... \int P(\theta_1, ..., \theta_p|Y) d\theta_2, ..., d\theta_p$$

Can use Monte Carlo sampling to estimate these integrals.

Need to confirm the above statement

3.8 Bayesian One Sample t-test

Likelihood: $Y_i|\mu, \sigma \sim N(\mu, \sigma^2)$ idep over $i = 1, ..., n$ Priors: $\mu \sim N(\mu_0, \sigma_0^2)$ independent of $\sigma^2 \sim InvGamma(a, b)$

Typically we are interested in marginal posterior because it accounts for uncertainty about σ^2

Marginal Posterior: $f(\mu|Y) = \int_0^\infty P(\mu, \sigma^2|Y) d\sigma^2,\ Y = (Y_1, ..., Y_n)$

if σ is known, the posterior of $\mu|Y$ is Gaussian and 95% Credible Interval is $E(\mu|Y) \pm Z_{0.975} SD(\mu|Y)$

if σ is unknown, the marginal (over σ^2) posterior of μ is t with $\nu = n+2a$ degrees of freedom.

$$E(\mu|Y) \pm t_{0.975}\ SD(\mu|Y)$$

$SD(\mu|Y)$: Standard Deviation

Can summarize results best in a table with Posterior Mean, Posterior SD, and 95% Credible Set.

3.9 Frequentist vs Bayesian Analysis of a Normal Mean

Frequentist

Estimate of the μ is $\bar{Y}$ If σ is known, the 95% C.I. is: $\bar{Y} \pm z_{0.975}\frac{\sigma}{\sqrt{n}}$

If σ is unknown, the 95% C.I. is: $\bar{Y} \pm t_{0.975,n-1}\frac{s}{\sqrt{n}}$

where t is the quantile of a t-distribution.

Bayesian

Estimate of μ is its marginal posterior mean.

Interval estimate is 95% Credible Interval.

If σ is known, Posterior of $\mu|Y$ is Gaussian

$E(\mu|Y) \pm Z_{0.975}SD(\mu|Y)$

If σ is unknown, the marginal (over σ^2) posterior of μ is t with $\nu = n+2a$ degrees of freedom.

$E(\mu|Y) \pm t_{0.975,\nu}\ SD(\mu|Y)$

3.10 Multiple Parameters in Multivariate Posteriors

Want to compute $P(\theta_2 > \theta_1|Y_1, Y_2)$.

Monte Carlo sampling of the posteriors a key tool!

Model is:

$$\begin{aligned} Y_1|\theta_1 &\sim Bin(N, \theta_1) \\ Y_2|\theta_2 &\sim Bin(N, \theta_2) \\ \theta_1, \theta_2 &\sim Beta(1, 1) \end{aligned} \tag{9}$$

Marginal Posteriors both independent of each other.

- $\theta_1|Y_1, Y_2 \sim Beta(Y_1 + 1, N - Y_1 + 1)$
- $\theta_2|Y_1, Y_2 \sim Beta(Y_2 + 1, N - Y_2 + 1)$

N <- 10; Y1 <- 5; Y2 <- 8;

S <- 10000

theta1 <- rbeta(S, Y1 + 1, N - Y1 + 1) theta2 <- rbeta(S, Y2 + 1, N - Y2 + 1)

(Y1 + 1) / (N + 2)

mean(theta1)

mean(theta2 > theta1)

3.11 Types of Uncertainty

Sampling

Parametric: Uncertainty about my guesses of the distribution of the parameter

3.11.1 Resolving Uncertainty

1. Plugin approach

   If $\hat{\theta}$ is an estimate, thus $Y^* \sim f(Y|\hat{\theta})$

   For example, Let $\hat{\theta} = \frac{2}{10}$. Predict $P(Y > 0) = 1 - (1 - 0.2)^{10}$.

   If $\hat{\theta}$ has small uncertainty, this is fine. Otherwise, this underestimated uncertainty in Y^*

2. Posterior Predictive Distribution (PPD)

   For the sake of prediction, the parameters aren't of interest as the parameters are vehicles by which the data inform about the predictive model.

   PPD averages over their posterior uncertainty which *accounts* for parametric uncertainty.

   $$f(Y^*|Y) = \int f(Y^*|\theta)p(\theta|Y)\ d\theta$$

   Input `= data Output =` prediction distribution

   > Given I've observed a certain amount of data Y, what is the distribution of the predictor values?

   Monte Carlo sampling approximates the PPD.

   (a) Example

   Let $\theta^{(1)}, ..., \theta^{(S)}$ be samples from the posterior.
   Let $Y^{*(s)} \sim f(Y|\theta^{(s)})$ where $Y^{*(s)}$ are samples from the PPD for each $\theta^{(s)}$
   Posterior Predictive Mean $\approx$ sample mean of the $Y^{*(s)}$
   $P(Y^* > 0) \approx$ sample proportion of non-zero $Y^{*(s)}$
   Y < -2; n <- 10;
   A <- Y + 1; B <- N - Y + 1

1-dbinom(0,10,0.2)
theta <- rbeta(100000,A,B) Ystar <- rbinom(100000,10,theta)
mean(Ystar>0)

4 Conjugate and Objective Priors (2020/10/01)

How do we choose priors? This is the most important step of a Bayesian Analysis.

Key Terms

- Conjugate vs Non-conjugate
- Informative vs Uninformative
- Proper vs Improper
- Subjective vs Objective

4.1 Conjugate

Def: Prior and Posterior Distribution are from the same parametric family. This is done through a pairing of the Likelihood Distribution and the Prior Distribution.

4.1.1 Beta-Binomial

Use Case: Estimating a Proportion!

- What is the probability of success for a new cancer treatment?
- What proportion of voters support a candidate?

Let $\theta \in [0,1]$ be a proportion we are trying to estimate.
Likelihood: $Y|\theta \sim Bin(n,\theta)$
Prior: $\theta \sim Beta(a,b)$
a: Prior number of successes b: Prior number of failures
Posterior: $\theta|Y \sim Beta(Y+a, n-Y+b)$

1. Frequentist Approach
   MLE: $\hat{\theta} = \frac{Y}{n}$
   $\hat{\theta} \sim N(\theta, \frac{\theta(1-\theta)}{n})$ for large Y and $n-Y$

Rule of Thumb for large enough n for proportions: At least 10-15 failures and 10-15 successes depending on which text book you read.

This is slightly different than the magic number 30 which is considered large enough for the mean.

$SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$

2. Proof

   (a) Short way

   The short proof uses "proportional to" ($\propto$) and hand waves the constants.

   Posterior:

$$\begin{aligned} f(\theta|Y) \propto & f(Y|\theta)f(\theta) = \binom{n}{Y}\theta^Y(1-\theta)^{n-Y} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ & \propto \theta^Y(1-\theta)^{n-Y}\theta^{a-1}(1-\theta)^{b-1} \\ & = \theta^{Y+a-1}(1-\theta)^{n-Y+b-1} \text{ (Looks like a Beta PDF)} \\ & \therefore \theta|Y \sim Beta(Y+a, n-Y+b) \end{aligned} \tag{10}$$

   (b) Long way

$$f(Y|\theta) = \frac{f(Y|\theta)\cdot f(\theta)}{f(Y)} = \frac{f(Y|\theta)\cdot f(\theta)}{\int_0^1 f(Y,\theta)d\theta} \tag{11}$$

   Numerator:

$$\begin{aligned} f(Y|\theta)f(\theta) = & \binom{n}{Y}\theta^Y(1-\theta)^{n-Y} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ = & \binom{n}{Y} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{Y+a-1}(1-\theta)^{n-Y+b-1} \end{aligned} \tag{12}$$

   Denominator:

$$
\begin{aligned}
f(Y) =& \int_0^1 \binom{n}{Y}\theta^Y(1-\theta)^{n-Y}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta \\
=& \binom{n}{Y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{Y+a-1}(1-\theta)^{n-Y+b-1}d\theta \\
=& \binom{n}{Y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(n+a+b)}\int_0^1 \frac{\Gamma(n+a+b)}{\Gamma(Y+a)\Gamma(n-Y+b)}\theta^{Y+a-1}(1-\theta)^{n-Y+} \\
\text{So? } f(y) =& \binom{n}{Y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(n+a+b)}
\end{aligned}
\tag{13}
$$

Posterior:

$$
\begin{aligned}
f(\theta|Y) =& \frac{\binom{n}{Y}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{Y+a-1}(1-\theta)^{n-Y+b-1}}{\binom{n}{Y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(n+a+b)}} \\
=& \frac{\Gamma(n+a+b)}{\Gamma(Y+a)\Gamma(n-Y+b)}\theta^{Y+a-1}(1-\theta)^{n-Y+b-1} \\
\theta \sim& Beta(Y+a, n-Y+b)
\end{aligned}
\tag{14}
$$

3. Shrinkage

   Posterior mean: $\hat{\theta}_B = E(\theta|Y) = \frac{Y+a}{n+a+b}$

   Posterior mean is between the sample proportion ($\frac{Y}{n}$) and the prior mean: $\frac{a}{a+b}$

   $\hat{\theta}_B = w\frac{Y}{n} + (1-w)\frac{a}{a+b}$

   where $w = \frac{n}{n+a+b}$

   - When n is large, $\hat{\theta}_B$ is closer to $\frac{Y}{n}$.
   - as a and b grow, posterior mean more dependent on the prior.

   Definition: The gravitation between the Likelihood function and the prior data. If there is

   What prior to select if research show θ is between 0.6 and 0.8? a = 7, b = 3 because $\frac{7}{7+3} = 0.7$

4.1.2 Related Problem using NegBin

Estimate the number of successes (Y) before n failures.

θ: probability of success

$\theta \sim Beta(a, b)$

$Y|\theta \sim NegBin(n, \theta)$

$$\begin{aligned} f(\theta|Y) \propto & \binom{Y+n+1}{Y} \theta^Y (1-\theta)^n \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \\ \propto & \theta^Y (1-\theta)^n \theta^{a-1}(1-\theta)^{b-1} \\ = & \theta^{Y+a-1}(1-\theta)^{n+b-1} \\ \sim & Beta(Y+a, n+b) \end{aligned} \tag{15}$$

4.1.3 Poisson-Gamma: One observation

Goal: Estimate a rate!

Let $\lambda > 0$ be the rate to be estimated.

- Observations made over a period of N and observe $Y \in \{0, 1, 2, ...\}$ events
- expected number of events: $N\lambda$

$\hat{\lambda} = \frac{Y}{n} = MLE$

Likelihood: $Y|\lambda \sim Poisson(N\lambda)$

Prior: $\lambda \sim Gamma(a, b)$

> λ is continuous and positive so Gamma is a natural distribution to use for estimating the rate.

Posterior: $\lambda|Y \sim Gamma(a + Y, b + N)$

Interpretation

a: Prior number of events b: Prior observation time

1. Proof (Short Way)

$$\begin{aligned} f(\lambda|Y) \propto & \frac{e^{-N\lambda}(N\lambda)^Y}{Y!} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \\ \propto & e^{-N\lambda} \lambda^Y \lambda^{a-1} e^{-b\lambda} \\ \propto & e^{-(N+b)\lambda} \lambda^{Y+a-1} \text{ (Looks like a Gamma)} \\ \therefore & \lambda|Y \sim Gamma(Y+a, N+b) \end{aligned} \tag{16}$$

2. Shrinkage

   The posterior mean is between the sample rate ($fracYN$) and the prior mean ($fracab$)

   $\hat{\lambda}_b = E(\lambda|Y) = \frac{Y+a}{N+b}$

   $\hat{\lambda}_B = w\frac{Y}{N} + (1-w)\frac{Y+a}{N+b}$

   where $w = \frac{N}{N+b}$

   What if we have no information about λ? In general, make PDF Wide

   What if λ is likely between 0.6 and 0.8? $a = 7$, $b = 10$ because $E(Y) = \frac{a}{b} = \frac{7}{10} = 0.7$

4.1.4 Poisson-Gamma: Two Observations

Likelihood:

$$\begin{aligned} f(Y_1, Y_2|\lambda) =& f(Y_1|\lambda) \cdot f(Y_2|\lambda) \text{ (if Y's are independent)} \\ =& \frac{(N\lambda)^{Y_1} e^{-N\lambda}}{Y_1!} \cdot \frac{(N\lambda)^{Y_2} e^{-N\lambda}}{Y_2!} \\ =& \frac{(N\lambda)^{Y_1+Y_2} e^{-2N\lambda}}{Y_1! \cdot Y_2!} \end{aligned} \tag{17}$$

Prior: $f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$
Posterior:

$$\begin{aligned} f(\lambda|Y_1, Y_2) \propto& \frac{(N\lambda)^{Y_1+Y_2} e^{-2N\lambda}}{Y_1! Y_2!} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \\ \propto& \lambda^{Y_1+Y_2} e^{-2N\lambda} \lambda^{a-1} e^{-b\lambda} \\ =& \lambda^{Y_1+Y_2+a-1} e^{-2N\lambda-b\lambda} \text{ (Looks like a Gamma)} \\ \therefore& \lambda|Y_1, Y_2 \sim Gamma(Y_1 + Y_2 + a, 2N + b) \end{aligned} \tag{18}$$

4.1.5 Poisson-Gamma: m Observations

Likelihood:

$$
\begin{aligned}
f(Y_1,...,Y_m|\lambda) =& f(Y_1|\lambda)\cdot ... \cdot f(Y_m|\lambda) \text{ (if Y's are independent)} \\
=& \Pi_1^m \frac{(N\lambda)^{Y_i} e^{-N\lambda}}{Y_i} \\
=& \frac{(N\lambda)^{\Sigma Y_i} e^{-mN\lambda}}{\Pi_m^1 Y_i!}
\end{aligned} \tag{19}
$$

Prior: $f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$
Posterior:

$$
\begin{aligned}
f(\lambda|Y_1,...,Y_m) \propto& \frac{(N\lambda)^{\Sigma Y_i} e^{-2mN\lambda}}{\Pi_{m\,Y_i!}^1 \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}} \\
\propto& (N\lambda)^{\Sigma Y_i} e^{-mN\lambda} \lambda^{a-1} e^{-b\lambda} \\
\propto& (N\lambda)^{a-1+\Sigma Y_i} e^{-(mN+b)\lambda} \text{ (Looks like a Gamma PDF)} \\
& \therefore \lambda|Y_1,...,Y_m \sim Gamma(\sum_1^m Y_i + a, mN + b)
\end{aligned} \tag{20}
$$

4.1.6 Gaussian-Gaussian

Goal: Estimate a mean! (μ)
Likelihood: $f(Y_1,...,Y_n|\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} exp(-\frac{1}{2\sigma^2} \sum_1^n (Y_i - \mu)^2)$
Prior:

$$
\begin{aligned}
f(\mu) =& \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{m}}} exp(-\frac{1}{2\frac{\sigma^2}{m}}(\mu - \theta)^2) \\
=& \frac{\sqrt{m}}{\sqrt{2\pi}\sigma}\ exp(-\frac{m}{2\sigma^2}(\mu - \theta)^2)
\end{aligned} \tag{21}
$$

Posterior:

$$
\begin{aligned}
f(\mu|Y_1,...,Y_n) \propto & \frac{1}{\sigma^n(2\pi)^{n/2}} exp(-\frac{1}{2\sigma^2}\Sigma(Y_i-\mu)^2) \cdot \frac{\sqrt{m}}{\sqrt{2\pi}\sigma} \\
& \propto exp(-\frac{1}{2\sigma^2}[\Sigma(Y_i-\mu)^2 + m(\mu-\theta)^2]) \\
& \propto exp(-\frac{1}{2\sigma^2}[\Sigma(Y_i^2 - 2\mu Y_i + \mu^2) + m(\mu^2 - 2\mu\theta + \theta^2)]) \\
& \propto exp(-\frac{1}{2\sigma^2}[2\mu\Sigma Y_i + n\mu^2 + m\mu^2 - 2m\mu\theta]) \text{ (where does the square Yi go?)} \\
& \propto exp(-\frac{1}{2\sigma^2}[2\mu n\bar{Y} + n\mu^2 + m\mu^2 - 2m\mu\theta]) \\
& = exp(-\frac{n+m}{2\sigma^2}[-2\frac{n\bar{Y}+m\theta}{n+m} + \mu^2]) \\
& \propto exp(-\frac{n+m}{2\frac{\sigma^2}{n+m}}[\mu - \frac{n\bar{Y}+m\theta}{n+m}]^2) \text{ (Looks like a Normal PDF)} \\
& \therefore \mu|Y_1,...,Y_n \sim N(\frac{n\bar{Y}+m\theta}{n+m}, \frac{\sigma^2}{n+m})
\end{aligned}
\tag{22}
$$

This can also be written as
Let $w = \frac{n}{n+m}$, then $\mu|Y_1,...,Y_m \sim N(w\bar{Y} + (1-w)\theta, \frac{\sigma^2}{n+m})$
m can *loosely* be interpreted as the prior number of observations

1. Shrinkage

   $\hat{\mu}_B = E(\mu|Y_1,...,Y_n) = w\bar{Y} + (1-w)\theta$ where $w = \frac{n}{n+m}$

   If no prior information available, make m small to make the prior uninformative. This is because a small m makes the variance large which makes the bell curve wide.

4.1.7 Gaussian-Gaussian: Known μ

If μ is known, then we should be estimating σ^2.

$\sigma^2 \sim Gamma(a,b)$ since Gamma is continuous over $(0,\infty)$ which matches the domain of the variance.

The math is easier if using a gamma prior for the inverse variance (τ).
Inverse Variance is also known as *precision* $\frac{1}{\sigma^2}$
If $\frac{1}{\sigma^2} \sim Gamma(a,b)$, then $\sigma^2 \sim InvGamma(a,b)$
Likelihood: $f(Y_1,...,Y_n|\sigma^2) = \frac{1}{(\sqrt{2\pi})^n(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}\sum_n^1(Y_i-\mu)^2)$

Prior: $f(\sigma^2) = \frac{b^a(\sigma^2)^{-a-1}e^{-b/\sigma^2}}{\Gamma(a)}$
Posterior:

$$\begin{aligned} f(\sigma^2|Y_1,...,Y_n) \propto & \frac{1}{\sqrt{2\pi}^n(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}\sum_1^n (Y_i-\mu)^2) \cdot \frac{b^a(\sigma^2)^{-b-1}}{\Gamma(a)} exp(\frac{-b}{\sigma^2}) \\ & \propto (\sigma^2)^{-n/2} exp(\frac{-1}{2\sigma^2}\sum_1^n (Y_i-\mu)^2) \\ & \propto (\sigma^2)^{-(\frac{n}{2}+a)-1} exp(\frac{-1}{\sigma^2}[\frac{\sum_1^n (Y_i-\mu)^2}{2} + b]) \end{aligned} \tag{23}$$

if μ is known, then $SSE = \sum_1^n (Y-\mu)^2$

$\sigma^2|Y_1,...,Y_n \sim InvGamma(\frac{n}{2}+a, \frac{SSE}{2}+b)$
Using τ
If Y_i has mean μ and precision τ, then likelihood is proportional to:

$$\Pi_n^1 f(y_i|\mu) \propto \ \tau^{n/2} exp(-\frac{\tau}{2}\sum_1^n (y_i-\mu)^2)$$

If $\tau \sim Gamma(a,b)$, then $\tau|Y \sim Gamma(\frac{n}{2}+a, \frac{SSE}{2}+b)$
This matches the results when using an Envisages for Variance.

1. Shrinkage

   Mean of InvGamma only exists for $a > 1$.

   Prior mean: $\frac{b}{a-1}$

   Posterior mean: $\frac{SSE+b}{n+2a-2}$

   common to make a,b small to give an uninformative prior. Then posterior mean converges towards sample variance.

4.2 Informative vs Uninformative

Can use informative priors from literature reviews, pilot studies, expert opinions, etc.

Prior Elicitation: Process of converting expert information *into* a prior. Experts may not know what an InvGamma is but their information can be converted into one!

Weak/Uninformative Priors commonplace. Easier to defend.

Strong Priors typically used for nuisance parameters. i.e. parameters we don't care about. The idea being that we don't care about it and really only want to affect the analysis if its *really* strong.

Sensitivity Analysis used to compare the posterior against several priors. This lets readers know how the prior exactly affects the analysis.

4.2.1 Mixture of Experts

Combine several priors into a single prior. For example, there are three studies which promote three different priors. These can be combined into a single prior.

$$\pi(\theta) = \sum_{j=1}^{J} w_j \pi_j(\theta)$$

where w_j is a weight with the constraints $w_j > 0$ and $\Sigma w_j = 1$

4.3 Improper Priors

A prior that doesn't integrate to 1. e.g. $\pi(\mu) = 1 \forall \mu \in \mathbb{R}$

It is okay to use an improper prior as long as you verify the posterior integrates to 1.

4.4 Subjective vs Objective Bayes

An objective analysis requires no subjective decisions by the analyst such as picking a prior, picking a likelihood function, treatment of outliers or transforms, etc.

Objective analysis may be feasible in a tightly controlled study but generally impossible for most analysis.

4.4.1 Objective Bayes

Lets an algorithm choose prior.

Examples

- Jeffrey's Prior
- Probability matching
- Maximum Entropy
- Empirical Bayes

- Penalized Complexity

Jeffrey's priors are most common.

Most of these are *improper* so posterior needs to be checked that it integrates to 1.

1. Jeffrey's Prior

   Jeffrey's Prior for θ: $p(\theta) = \sqrt{I(\theta)}$ where $I(\theta)$ is the Fisher Information Matrix.

   $$I(\theta) = -E_{Y|\theta}[\frac{d^2}{d\theta^2} logp(Y|\theta)]$$

   Once the likelihood is specified, Jefferey's prior is determined with no additional input hence being objective about prior.

   Examples

   Likelihood: $Y \sim Bin(n, \theta)$

   Jefferey's Prior: $\theta \sim Beta(0.5, 0.5)$

   Likelihood: $Y \sim N(\mu, 1)$

   Jefferey's Prior: $p(\mu) \propto 1$

   Likelihood: $Y \sim N(0, \sigma^2)$

   Jefferey's Prior: $p(\sigma) \propto 1/\sigma$

2. Reference Priors

   Try to be uninformative. Univariate models give Jeffreys Priors. Multivariate models give different priors. Harder to compute than Jeffrey's.

3. Probability Matching Priors (PMP)

   Designed so Posterior Credible Intervals have correct frequentist coverage.

   For example, if $Y_i|\mu \sim N(\mu, 1)$, the PMP is $p(\mu) = 1$. Then, Posterior is $\mu|Y \sim N(\bar{Y}, 1/n)$

   Only a few cases where this can be used.

4. Empirical Bayes

   Pick priors based on data.

   Ex: Maybe σ^2 has prior mean s^2

   Criticized for using data twice: once for the prior, and once for the likelihood.

5. Penalized Complexity Priors (PCP)

   A PCP prior begins with a simple base model. e.g linear regression with all slopes equal to 0.

   Full model is shrunk towards base model. e.g regression with non-zero slopes.

   Distance from full to base model has exponential prior to penalize the more complex model from deviating from the base.

   Requires picking the parameter in the exponential prior and setting priors for the parameters in the base model **so not purely objective**

6. Maximum Entropy Priors

   Entropy is a measure of uncertainty. The entropy of the PMF $f(x)$ is

   $$-\Sigma_{x \in S} f(x) log[f(x)]$$

   (a) Fix a few quantities of the prior distribution. e.g. $E(\theta) = 0.5$

   (b) Find the prior with maximum entropy that satisfies these constraints.

   If θ has support $\mathbb{R}$ and mean and variance are known, maximum entropy prior is Gaussian.

   Not purely objective because you have to set the constraints

   *

5 Deterministic Methods & MCMC Sampling (2020/10/08)

The big question is **How to summarize the Posterior**?

We need point estimates, credible sets, etc.

Algorithms to Estimate Complicated Joint Posteriors

- Use a point estimate (e.g. MAP), ignore uncertainty
- Approximate Posterior as Gaussian using Bayesian Central Limit Theorem
- Numerical Integration. Not touched on much here. Moreso in Numerical Analysis
- Markov-Chain Monte Carlo Sampling

5.1 MAP Estimation (Maximum a Posteriori)

Sometimes you don't need an entire posterior distribution. A single point estimate will do. For example, prediction in Machine Learning.

MAP Estimate is the posterior **mode**. AKA the peak of the posterior distribution.

$$\hat{\theta}_{MAP} = \underset{\theta}{argmax}\ p(\theta|Y) = \underset{\theta}{argmax}\ log[f(Y|\theta)] + log(\pi(\theta))$$

Similar to MLE but includes prior.

5.1.1 Example

Let $Y|\theta \sim Gamma(Y+a, N+b)$ and $\theta \sim Gamma(a,b)$. Find $\hat{\theta}_{MAP}$

$$p(\theta|Y) = \frac{(N+b)}{Y+a}\Gamma(Y+a)\theta^{Y+a-1}\ e^{-(N+b)\theta}$$

$$\begin{aligned}
\hat{\theta}_{MAP} =& \underset{\theta}{argmax}\ log[f(\theta|Y)] \\
=& \underset{\theta}{argmax}\{(Y+a)log(N+b) - log(\Gamma(Y+a)) + (Y+a-1)log(\theta) - (N+b)\theta\} \\
=& \frac{d}{d\theta}\ log(p(\theta|Y)) = \frac{y+a-1}{\theta} - (N+b) = 0 \\
=& \frac{Y+a-1}{N+b}
\end{aligned} \tag{24}$$

5.2 Bayesian Central Limit Theorem

Berstein-Von Mises Theorem: As the sample size grows, the posterior doesn't depend on the prior. i.e. Shrinkage.

Def: For large N and some other conditions, $\theta|Y \approx Normal$

$$\theta|Y \sim N(\hat{\theta}_{MAP}, I(\hat{\theta}_{MAP})^{-1})$$

I is Fisher's Information Matrix, aka the Hessian Matrix.

$$I_{jk} = \frac{-d^2}{d\theta_j\ d\theta_k} log[p(\theta|Y)]$$

evaluated at $\hat{\theta}_{MAP}$

5.2.1 Example

Let $\theta \sim Beta(0.5, 0.5)$ and $Y|\theta \sim Bin(n, \theta)$. Find Gaussian approximation for $p(\theta|Y)$.

In this case, $Beta(0.5, 0.5)$ is Jeffreys Prior.

Posterior: $\theta|Y \sim Beta(Y + 0.5, n - Y + 0.5)$
Need a MAP Estimator.

$$\frac{\Gamma(Y + 1 + n - Y)}{\Gamma(Y + 0.5)\Gamma(n - Y + 0.5)}\Theta^{Y-0.5}(1-\theta)^{n-Y-0.5}$$

$$\begin{aligned}
log p(\theta|Y) =& log\Gamma(n+1) - log\Gamma(Y+0.5) - log\Gamma(n-y+0.5) + (y-0.5)log\theta + (n-y-0.5)log\theta \\
\frac{d}{d\theta}\ log p(\theta|Y) =& \frac{Y-0.5}{\theta} - \frac{n-y-0.5}{1-\theta} = 0 \\
=>& \frac{Y-0.5}{\theta} = \frac{n-Y-0.5}{1-\theta} \\
=>& \hat{\theta}(n-Y-0.5) = (1-\hat{\theta})(Y-0.5) \\
=>& \hat{\theta}(n-Y-0.5) + \hat{\theta}(Y-0.5) = Y - 0.5 - \hat{\theta}(Y-0.5) \\
=>& \hat{\theta}(n-Y-0.5) = Y - 0.5 \\
=>& \hat{\theta}(n-Y-0.5+Y-0.5) = Y - 0.5 \\
=>& \hat{\theta}(n-1) = Y - 0.5 \\
\therefore& \hat{\theta}_{MAP} = \frac{Y-0.5}{n-1}
\end{aligned} \tag{25}$$

Finding the Variance via the Information Matrix

$$\begin{aligned}
\frac{-d^2}{d\theta^2} log\ p(\theta|Y) =& \frac{-d}{d\theta}[\frac{d}{d\theta} logp(\theta|Y)] \\
=& \frac{-d}{d\theta}[\frac{Y-0.5}{\theta} - \frac{n-Y-0.5}{1-\theta}] \\
=& -[-\frac{Y-0.5}{\theta^2} - \frac{n-Y-0.5}{(1-\theta)^2}] \\
=& \frac{Y-0.5}{\theta^2} + \frac{n-Y-0.5}{(1-\theta)^2} \\
=& \frac{Y-0.5}{\theta^2} + \frac{n-Y-0.5}{[1-(\frac{Y-0.5}{n-1})]^2}
\end{aligned} \tag{26}$$

$$
\begin{aligned}
I(\hat{\theta}_{MAP}) =& \frac{Y-0.5}{[\frac{Y-0.5}{n-1}]^2} + \frac{n-Y-0.5}{[1-\frac{Y-0.5}{n-1}]^2} \\
=& \frac{(n-1)^2}{Y-0.5} + \frac{n-Y-0.5}{[\frac{n-1}{n-1} - \frac{Y-0.5}{n-1}]^2} \\
=& \frac{(n-1)^2}{Y-0.5} + \frac{n-Y-0.5}{[\frac{n-Y-0.5}{n-1}]^2} \\
=& \frac{(n-1)^2}{Y-0.5} + \frac{(n-1)^2}{n-Y-0.5} \\
=& (n-1)^2 \frac{1}{Y-0.5} + \frac{1}{n-Y-0.5} \\
=& (n-1)^2 \frac{n-1}{(Y-0.5)(n-Y-0.5)} \\
=& \frac{(n-1)^3}{(Y-0.5)(n-Y-0.5)}
\end{aligned}
\tag{27}
$$

$$\therefore \theta|Y \approx N(\tfrac{Y-0.5}{n-1}, \tfrac{(Y-0.5)(n-Y-0.5)}{(n-1)^3})$$

Note that $I(\hat{\theta}_{MAP})$ produces the Inverse Variance hence why the Variance of the approximation is flipped.

Normal Approximation for Beta-Binomial using Bayesian Central Limit Theorem.

For large Datasets with small number of params, the Normal approximation is good.

5.3 Numerical Integration

Only feasible for small p.

Iteratively Nested Laplace Approximation (INLA) combines Gaussian approximation with numerical integration. It works well if most parameters are approximately normal.

5.4 Monte Carlo Sampling

Collection of all params in model: $\theta = (\theta_1, ..., \theta_p)$

Dataset: $Y = (Y_1, ..., Y_n)$

Posterior Distribution: $f(\theta|Y)$

If $\theta^{(1)}, ..., \theta^{(s)}$ are samples from $f(\theta|Y)$, then mean of the S samples approximate a posterior mean.

Most common MCMC Algorithms: Gibbs, Metropolis

5.4.1 Gibbs Sampling

- Sample from High dimension Posteriors
- Break problem of sampling from the high dimension joint distribution into a series of samples from low dimensional conditional distributions.

Samples are not independent. The dependencies form a Markov Chain.

Full Conditional (FC) Distribution: Distribution of one parameter taking all other parameters as *fixed and known*.

1. MCMC for Bayesian T-Test

   $Y_i \sim N(\mu, \sigma^2)$ where $\mu \sim N(0, \sigma_0^2)$ and $\sigma^2 \sim InvGamma(a, b)$

   FC Distribution 1

   $$\mu|\sigma^2, Y \sim N(\frac{n\bar{Y}\sigma^{-2} + \mu_0\sigma_0^{-2}}{n\sigma^{-2} + \sigma_0^{-2}}, \frac{1}{n\sigma^{-2} + \sigma_0^{-2}})$$

   FC Distribution 2

   $$\sigma^2|\mu, Y \sim InvGamma(\frac{n}{2} + a, \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^2 + b)$$

2. Algorithm

   (a) Set Initial values for all parameters: $\theta_1^{(0)}, ..., \theta_p^{(0)}$

   (b) Sample variables one at a time from Full Conditional: $p(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_p, Y)$

   - Rather than 1 p-dimensional samples, we produce p 1-dimensional samples
   - Repeat until required number of samples are generated.

   Example

   (a) Set initial values

   (b) For iteration t

- FC1: Draw $\theta_1^{(t)}|\theta_2^{(t-1)}, ..., \theta_p^{(t-1)}, Y$
- FC2: Draw $\theta_2^{(t)}|\theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_p^{(t-1)}, Y$
- ...
- FCp: Draw $\theta_p^{(t)}|\theta_1^{(t)}, ..., \theta_{p-1}^{(t)}, Y$

Repeat 2 $\mathbb{S}$ times giving posterior draws $\theta^{(1)}, ..., \theta^{(S)}$

Why does it work?

Theorem: For any initial values, the chain will eventually converge to the posterior.

Theorem: If $\theta^{(s)}$ is a sample from the posterior, then $\theta^{(s+1)}$ is too.

Once the chain has *converged*, then discard the first T samples as "burn in". Use remaining S - T to approximate the Posterior.

> I have also heard "burn in" described as annealing in my Monte Carlo class.

3. Practice Problem

Work out the full conditionals for λ and b for the following model:

$$\begin{aligned} Y|\lambda, b \sim & Poisson(\lambda) \\ \lambda|b \sim & Gamma(1, b) \\ b \sim & Gamma(1,1) \sim exp(1) \end{aligned} \tag{28}$$

Posterior:

$$\begin{aligned} P(\lambda, b|Y) \propto & f(Y|\lambda, b) f(\lambda|b) = f(Y|\lambda, b) f(b) f(\lambda|b) \\ = & \frac{e^{-\lambda}\lambda^Y}{Y!} \cdot e^{-b} \cdot b \; e^{-b\lambda} \\ \propto & e^{-\lambda}\lambda^Y e^{-b} b \; e^{-b\lambda} \\ = & b e^{-\lambda - b - b\lambda} \lambda^Y \end{aligned} \tag{29}$$

Full Conditional Distributions

(a) Write in terms of λ. Everything else is fixed.

$$\begin{aligned}\lambda|b, Y \propto & \lambda^{Y} e^{-\lambda} e^{-b\lambda} \\ = & \lambda^{(y+1)-1} e^{-(b+1)\lambda} \\ \text{So? } & \lambda|b, Y \sim Gamma(y+1, b+1)\end{aligned} \tag{30}$$

(a) Write in terms of b. Everything else is fixed.

$$\begin{aligned}b|\lambda, Y \propto & b e^{-b} e^{-b\lambda} = b\ e^{-(\lambda+1)b} \\ \text{So? } & b|\lambda, Y \sim Gamma(2, \lambda+1)\end{aligned} \tag{31}$$

6 MCMC Sampling & Convergence (2020/10/15)

Adaptive MCMC: Uses same candidate distribution throughout the chain but adjusts hyper-parameters to tune Acceptance Rate.

Hamiltonian MCMC: Uses gradient of the posterior to adjust hyper-parameters throughout the chain. Default MCMC method in STAN.

3 Main Decisions for MCMC Algo

1. Select Initial Values
   - Use Method of Moments or MLE
   - Pick purposefully bad values to demonstrate convergence.
2. Determine the chain convergence
3. Determine how many samples you need (Effective Sample Size)

The following MCMC Sampling methods can be mixed and matched as needed to use and sample from candidate distributions (barring Gibbs).

6.1 Metropolis-Hastings Sampling

Generic case which allows for Asymmetric Candidate Distributions.

Let θ_j^c be a Random Variable from the candidate distribution

$$\theta_j^c \sim q(\theta|\theta^*)$$

For example, if $\theta_j^c \sim N(\theta_j^*, s_j^2)$, then

$$q(\theta_j^c|\theta^*) = \frac{2}{s_j\sqrt{2\pi}} exp[-\frac{(\theta_j^c - \theta_j^*)^2}{2s_j^2}] = q(\theta^*|\theta_j^c)$$

High correlation cases cause slow convergence. If high correlation is present, **Blocked Gibbs/Metropolis** can help.

Ex. Linear Regression iterates between sampling the block $(\beta_1, ..., \beta_p)$ and σ^2

Blocked still not clear. Follow up with more.

6.2 Gibbs Sampling

Gibbs samples each parameter from its conditional distribution. Which makes use of conjugate priors. It is not obvious to use without conjugate priors.

$$P(\mu|Y) \propto exp[-\frac{1}{2}(Y - \mu)^2] \cdot \mu^{a-1}(1 - \mu)^{b-1}$$

Where $Y \sim N(\mu, 1), \mu \sim Beta(a, b)$

No known conjugate prior for some likelihoods. e.g. Logistic Regression.

This is where Metropolis Sampling comes in!

Special Case of Metropolis sampling where acceptance rate is always 1 and the proposal distributions are the posterior conditionals.

6.3 Metropolis Sampling

A version of rejection Sampling

Let θ_j^* be the current value of the parameter being updated and θ_j be the current value for all parameters.

Let θ_j^c be the candidate value where

$$\theta_j^c \sim N(\theta_j^*, s_j^2)$$

Let R be the probability of accepting a move where

$$R = min\{1, \frac{P(\theta_j^c|\theta_{(j)}, Y)}{P(\theta_j^*|\theta_{(j)}, Y)}\}$$

This is a special case of Metropolis-Hastings sampling where the candidate distribution is **symmetric**.

6.3.1 Tuning s_j

s_j is a hyper-parameter for the Metropolis Algorithm. Ideally, it is somewhere between 0.3 and 0.4. This is because we don't want to accept or reject *too* much.

If s_j is small

- proposed distribution is narrow
- Nearly all candidates are accepted

If s_j is large

- proposed distribution is wide
- nearly all candidates rejected. As in, there are a lot of straight lines on the trend graph.

6.3.2 Logistic Regression Example

This is how Bayesians see Logistic Regression.

$$Y_i|\theta \sim Bern(logit^{-1}(\theta)) = Bern(\frac{e^\theta}{1+e^\theta})$$

Where $\theta \sim N(\mu_0, \sigma_0^2)$

6.4 Convergence Diagnostics

6.4.1 Geweke

Compares mean at beginning of chain with mean at end of chain.

Test statistic: Z-score.

$|Z| > 2$ == poor convergence.

6.4.2 Gelman-Rubin

By running multiple chains, hopefully to see the same result. The measurements between chains should agree. Essentially an ANOVA test of whether the chains have the same mean.

1 is perfect 1.1 is decent but not great convergence.

6.4.3 Effective Sample Size (ESS)

Idea

Highly correlated samples have less information than independent samples. ESS accounts for this autocorrelation. For example, you may think you have 10,000 sampes but you actually have less than 10K *good* samples because you need to account for autocorrelation.

Ideally, samples independent across iterations but sometimes not the case. Lower values are better but if chains are long enough, then large values can be ok.

S: Number of MCMC Samples

$\rho(h)$: Autocorrelation with Lag h.

For example, $\rho(1)$ is autocorrelation coefficient with the first Lag.

$$ESS = \frac{S}{1 + 2\sum_{h=1}^{\infty} \rho(h)}$$

ESS should be at least a few **thousand**.

Naive SE: $SE = \frac{S}{\sqrt{S}}$

Time Series SE is more realistic.

$SE = \frac{S}{\sqrt{ESS}}$

Thinning: Use every K iterations to reduce autocorrelation.

6.5 Handling Massive Datasets

- MAP Estimate (Prediction only. Good for ML)
- Bayesian CLT
- Variation Bayes: Approximate Posterior by assuming posterior independent across all parameters. (Fast but questionable statistical properties).
- Parallel Computing
- Batch Datasets to Divide and Conquer

Misc thing about JAGS, STAN.

Sometimes people use a cauchy distribution to do a standard deviation in JAGS, STAN.

7 Bayesian Linear Modeling (2020/10/22)

7.1 Linear Regression

$$Y_i = \beta_0 + \sum_1^p \beta_i X_i + \epsilon_i,\ \epsilon_i \sim iidN(0, \sigma^2)$$

$Y_i \sim N(\beta_0 + \sum_1^p \beta_i X_i, \sigma^2)$

Bayesian and classical Linear Regression are similar if $n >> p$ and the priors are uninformative. Results can be different for challenging problems though.

$n >> p$: n is much, much larger than p

Metric-predicted Variable on 1-2 groups == One sample, Two Sample t-test.

7.1.1 Bayesian One-Sample t-test

$$Y_i, ..., Y_n \sim N(\mu, \sigma^2)$$

Typically Y_i is a difference (post - pre) and checking if average $\neq 0$

1. Known Sigma

   Under Jeffreys Prior, $\pi(\mu) = 1$ with fixed σ.

   $$\mu|Y, \sigma \sim N(\bar{Y}, \frac{\sigma^2}{n})$$

   Posterior Mean: $\bar{Y}$

   Credible Set: $\bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$

   $P(H_0|Y) = P(\mu \leq 0|Y) = \Phi\left(\frac{\sqrt{n}\bar{Y}}{\sigma}\right)$

2. Unknown Sigma

   Jeffreys Prior is $\pi(\mu, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{3/2}$

   Marginal Posterior integrating over uncertainty in σ^2 is $\mu|Y \sim t_n(\bar{Y}, \frac{\hat{\sigma}^2}{n})$

   where $\hat{\sigma}^2 = \frac{1}{n}\sum_1^n (y_i - \bar{Y})^2$

   > The main difference between Frequentist and Bayesian t-test is n-1 vs n respectively for the degrees of freedom for the t-distribution.

7.1.2 Bayesian Two-Sample t-test

n_1 observations from group 1: $Y_i \sim N(\mu, \sigma^2)$
n_2 observations from group 2: $Y_i \sim N(\mu + \delta, \sigma^2)$
Goal: Compare δ to 0.

1. Sigma Known

   Jeffreys prior: $\pi(\mu, \delta) = 1$

   Identical to the two-sample z-test

2. Sigma unknown

   Jeffreys prior: $\pi(\mu, \delta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^2$

   Marginal Posterior integrating uncertainty in σ^2 and μ

$$\delta | Y \sim t_n(\bar{Y}_2 - \bar{Y}_1, \frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2})$$

   where $\hat{\sigma}^2 = \frac{1}{n}\left[\sum_1^{n_1}(Y_i - \bar{Y}_1)^2 + \sum_1^{n_2}(Y_i - \bar{Y}_2)^2\right]$

   The main difference between Frequentist and Bayesian two-sample t-test is $n - 2$ compared to $n = n_1 + n_2$ respectively for the degrees of freedom for the t-distribution.

7.2 Review of Least Squares

For Simple Linear Regression

$$\begin{aligned}
\vec{\beta}_{OLS} =& \underset{\vec{\beta}}{argmin} \sum_{i=1}^{n} \left[(Y_i - [\beta_0 + \beta_1 X_i])^2\right], \text{ Considered Q(B0, B1)} \\
\frac{\partial}{\partial \beta_0} Q(\beta_0, \beta_1) =& \sum_{i=1}^{n} [-2(Y_i - [\beta_0 + \beta_1 X_i])] = 0 \\
\frac{\partial}{\partial \beta_1} Q(\beta_0, \beta_1) =& \sum_{i=1}^{n} [-2X_i(Y_i - [\beta_0 + \beta_1 X_i])] = 0
\end{aligned} \tag{32}$$

7.3 Bayesian Regression

$$Y_i \sim N(\beta_0 + \sum_1^p X_i\beta_i, \sigma^2)$$

Crucial to verify appropriateness with QQ Plots, Added Variable Plots, etc.

Since this is a **Bayesian** analysis, priors for β, σ are required.

For the purpose of setting priors, it is helpful to standardize both the response and the covariates to have mean 0 and std. dev 1.

Many priors have been considered:

1. Improper
2. Gaussian. Make it uninformative.
3. Double Exponential Priors. Bayesian version of Lasso
4. Many more

7.3.1 Improper Priors

1. Known Sigma

   With σ fixed, Jeffreys prior is flat. $P(\beta) = 1$.

   $$\beta|Y \sim N(\hat{\beta}_{OLS}, \sigma^2(X^TX)^{-1})$$

   Likelihood: $\vec{Y} \sim N(X\vec{\beta}, \sigma^2 I)$

   Prior: $P(\vec{\beta}) = 1$

   Posterior:

   $$\begin{aligned} P(\vec{\beta}|\vec{Y}) &\propto P(\vec{Y}|\vec{\beta})P(\vec{\beta}) \\ &\propto exp(-\frac{1}{2\sigma^2}(\vec{Y} - X\vec{\beta})^T(\vec{Y} - X\vec{\beta})) \text{Kernel of Multivariate Normal} \end{aligned} \tag{33}$$

   Recall: $\hat{\beta} = (X^TX)^{-1}X^TY$

$$
\begin{aligned}
(\vec{Y}-X\vec{\beta})^T(\vec{Y}-X\vec{\beta}) =&(\vec{Y}^T-\vec{\beta}^TX^T)(\vec{Y}-X\vec{\beta})\\
=&Y^TY-Y^TX\beta-\beta^TX^TY+\beta^TX^TX\beta\\
=&Y^TY-\underbrace{Y^TX(X^TX)^{-1}}_{\hat{\beta}^T}(X^TX)\beta-\beta^T(X^TX)\underbrace{(X^TX)^{-1}X^TY}_{\hat{\beta}}+\beta^TX^TX\beta\\
=&Y^TY-\hat{\beta}^T(X^TX)\beta-\beta^T(X^TX)\hat{\beta}+\hat{\beta}^T(X^TX)\beta\\
=&Y^TY-\hat{\beta}^T(X^TX)\hat{\beta}-\hat{\beta}^T(X^TX)\beta-\beta^T(X^TX)\hat{\beta}+\beta^T(X^TX)\beta+\hat{\beta}^T(X^T\\
=&Y^TY-\hat{\beta}^T(X^TX)\hat{\beta}+\beta^T(X^TX)\beta-\beta^T(X^TX)\hat{\beta}-\beta^T(X^TX)\beta+\hat{\beta}^T(X^T\\
=&Y^TY-\hat{\beta}^T(X^TX)\hat{\beta}+(\beta^T-\hat{\beta}^T)(X^TX)(\beta-\hat{\beta})\\
=&Y^TY-\hat{\beta}^T(X^TX)\hat{\beta}+(\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta})
\end{aligned}
\tag{34}
$$

Posterior:

$$
\begin{aligned}
P(\vec{\beta}|\vec{Y}) \propto& exp\left(-\frac{1}{2\sigma^2}\left[Y^TY-\hat{\beta}^T(X^TX)\hat{\beta}+(\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta})\right]\right)\\
\propto& exp\left(-\frac{1}{2\sigma^2}\left[(\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta})\right]\right)\\
\therefore& \vec{\beta}|Y\sim N(\hat{\beta},\sigma^2(X^TX)^{-1})
\end{aligned}
\tag{35}
$$

2. Unknown Sigma

   Sigma is rarely known so $\sigma^2 \sim InvGamma(a,b)$ where $a=b=0.1$ so that the prior is uninformative.

   The Posterior of β is a Multivariate t-distribution centered on β.

   Jeffreys Prior: $P(\beta,\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{P}{2}+1} \sim InvGamma(P/2,0)$

   Posterior: $\beta|Y \sim t_n(\hat{\beta}_{OLS},\hat{\sigma}^2(X^TX)^{-1})$

   where $\hat{\sigma}^2=(Y-X\hat{\beta}_{OLS})^T(Y-\hat{\beta}_{OLS})/n$

   The posterior is proper if the Least Squares solution exists.

7.3.2 Gaussian Priors

1. Multivariate Gaussian Prior

Zellner's g-prior: $\beta \sim N(\vec{0}, \frac{\sigma^2}{g}(X^T X)^{-1})$. Shrinkage of variance by a factor of g.

Posterior mean: $\frac{1}{1+g}\hat{\beta}_{OLS}$

g shrinks Least Squares estimate towards 0.

$g = \frac{1}{n}$ is common. It is called the *Unit Information* Prior.

2. Univariate Gaussian Priors

   If there are many covariates or the covariates are collinear, then $\hat{\beta}_{OLS}$ is unstable.

   Independent priors can counteract collinearity.

$$\beta_j \sim N(0, \frac{\sigma^2}{g})$$

   independent over j

   Posterior Mode: $\underset{\beta}{argmin} \sum_1^n (Y_i - \mu_i)^2 + g \sum_1^p \beta_j^2$

   This is known as **Ridge Regression** in Frequentist Statistics and is used to stabilize the Least Squares solution.

7.3.3 Bayesian Lasso (BLASSO)

Prior: $\beta_j \sim DE(\tau)$

$f(\beta) \propto exp\left(-\frac{|\beta|}{\tau}\right)$

- Square in the Gaussian Prior is replaced with an absolute value.
- Shape more peaked at 0
- Favors settings where many β_j are near zero and a few large β_j
- p is large but most covariates are noise

Posterior mode: $\underset{\beta}{argmin} \sum_1^n (Y_i - \mu_i)^2 + g \sum_1^p |\beta_j|$

Adds stability by shrinking estimates towards zero, sets some coefficients to zero.

Performs variable selection **and** estimation making it convenient.

Sampling

- Full conditionals don't exist in closed form but there's a trick to make Gibbs Sampling work.
- Metropolis sampling works

Can use 'BLR' R package.

7.4 Summarization

Table with marginal means and 95% Credible Intervals useful but unwieldy for large p.

Picking a subset of covariates is very important.

Cross-validation, stochastic searches, and information critera can be used to vet covariates.

7.5 Predictions

Given a new covariate vector X_{new}, we would like to predict the response Y_{new}.

Cannot just "plug-in" $\hat{\beta},\ \hat{\sigma}$ for $\beta,\ \sigma^2$ in

$$Y_{new}|\hat{\beta},\hat{\sigma} \sim N(X_{new}\hat{\beta},\hat{\sigma}^2)$$

The prediction intervals will be too narrow. Also, Uncertainty about $\beta,\ \sigma$ needs to be accounted for when making predictions.

7.5.1 Posterior Predictive Distribution (PPD)

$$\begin{aligned} P(Y_{new}|Y) &= \int f(Y_{new},\beta,\sigma|Y)d\beta d\sigma \\ &= \int f(Y_{new}|\beta,\sigma)f(\beta,\sigma|Y)d\beta d\sigma \end{aligned} \tag{36}$$

Marginalizing over the model parameters accounts for their uncertainty.

The concept of PPD applies generally (e.g. logistic regression) and means the distribution of the predicted value marginally over model parameters.

MCMC naturally draws from Y_{new}'s PPD.

For a given MCMC iteration t:

- $\beta^{(t)},\ \sigma^{(t)}$ are samples
- $Y_{new}^{(1)},...,_{new}^{(S)}$ are sampled from the PPD

$$Y_{new}^{(t)} \sim N(X\beta^{(t)}, \sigma^{(t)2})$$

This is an example of the claim that "Bayesian methods naturally quantify uncertainty."

8 Advanced Modeling (2020/10/29)

Advanced Modeling meaning not just Linear models.

8.1 GLMs (Logistic)

$$logit = log\left(\frac{P(Y=1)}{P(Y=0)}\right) = log\left(\frac{X}{1-X}\right) \text{ log-odds}$$

β_j represents the increase in the log-odds of an event corresponding to a one-unit increase in the covariate.

$$logit\left[P(Y_i=1)\right] = \eta_i = \beta_0 + \sum_1^p \beta_{ij} X_{i1}$$

The inverse of the logit transformation is the expit

$$expit(x) = \frac{exp(x)}{exp(x)+1}$$

$$P(Y_i = 1) = expit(\eta_i) \in [0,1]$$

8.1.1 Logistic Regression - Bayesian

Logistic Regression requires a prior for β. Zellner, BLASSO, etc. can be used. Full conditional distributions are not conjugate so Metropolis Sampling must be used.

8.1.2 Beta Regression - Bayesian

Can be used when the response takes on values between 0 and 1.

8.2 Random Effects

Linear Regression assumes error terms are independent; however, this is violated if there are observations that are grouped. e.g. multiple classrooms of students may be independent but students within a class may not be.

Random effects are a natural way to account for this independence.

> See GLM notes. Much of this was covered in Generalized Linear Models

8.2.1 One-way Random Effects Models

Let the score for student i in class j be represented as:

$Y_{ij} = \alpha_j + \epsilon_{ij}$

Random effect for classroom j is a considered a random draw from a population: $\alpha_j \sim N(\mu, \tau^2)$

$\epsilon_{ij} \sim N(0, \sigma^2)$: irreducible random error for each observation.

1. Bayesian Modeling

   μ, σ^2, τ are parameters and thus must have priors.

   $\mu \sim N()$ with some Large variance

   $\sigma^2, \tau^2 \sim InvGamma(a, b)$

   Full Conditionals are conjugate making MCMC Sampling very fast **BUT** under the Inverse Gamma Prior for the variances, the induced priors for σ, τ have no mass at zero. e.g. the density is not flat around zero.

   The **Half-Cauchy Prior** is recommended for the Standard Deviation to rectify this.

   $$P(\sigma) = \frac{2}{\pi(1 = \sigma^2)}$$

   The Half-Cauchy prior is essentially a Student-t distribution with 1 degree of freedom restricted to be positive. This is flat around zero with heavy tails which is desirable for a prior.

   Examples will likely give similar results but it is preferred to use Half-Cauchy.

   MCMC treats Random Effects like other parameters. i.e. all parameters are random. But it's still considered a Random Effect from the Model perspective.

8.3 Linear Mixed Models (LMM)

Linear Model *with* Random Effects.

$$Y_{ij} = \beta_0 + X_{ij}\beta_1 + \alpha_j + \epsilon_{ij}$$

X_{ij}: Age of student i in class j

Regression coefficients β_0, β_1 that apply to all students are called *fixed effects*. α_j is the random effect.

8.3.1 Random Slopes Models

Let Y_{ij} be the j-th observation for subject i.

Sometimes specifying a different regression for each subject to capture variability over a population makes sense.

1. Example

   Let Y_{ij} be the bone density for child i and at X_j

   Bone Density can be modeled for a different regression for each child to capture the variability over a population of children.

   $$Y_{ij} \sim N(\gamma_{0i} + X_i\gamma_{1i}, \sigma^2)$$

   $\gamma_i = (\gamma_{i0}, \gamma_{i1})^T$ controls the growth curve for child i.

   The prior $\gamma_i \sim N(\beta, \Sigma)$ ties all the separate regressions together. γ_i are the random effects specific to one child. β are fixed effects common to all children.

   The hierarchical Model is:

   - $Y_{ij} \sim N(\gamma_{0i} + X_i\gamma_{i1}, \sigma^2)$
   - $\gamma \sim N(\beta, \Sigma)$
   - $p(\beta) \propto 1$
   - $\sigma^2 \sim InvGamma(a, b)$
   - $\Sigma \sim InvW(k, R)$

   The full conditionals are all conjugate!

2. Random Effects Covariance Matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Using the above example,

σ_1^2: Variance of the **intercepts** across children

σ_2^2: Variance of the **slopes** across children

σ_{12}: Covariance between intercepts and slopes

Prior 1

$\sigma_1^2, \sigma_2^2 \sim InvGamma(a, b), \rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \sim Uni(-1, 1)$

Works better for smaller dimensions.

Prior 2

Inverse Wishart is most common prior for a $p \times p$ covariance matrix. Works better in higher dimensions.

Reduces to Inverse Gamma if $p = 1$.

Let $\Sigma \sim InvW(k, R),\ k > p + 1$ and R is a $p \times p$ covariance matrix. k, R are hyper-parameters.

PDF:

$$f(\Sigma) \propto |\Sigma|^{-(k+p+1)/2} exp\left[\frac{1}{2} trace(R\Sigma^{-1})\right]$$

$E(\Sigma) = \frac{1}{k-p-1} R$

8.4 Linear Models with Correlated Errors

One can model correlated errors directly as an alternative to Random Effects.

Let Y_i be some measurement at s_i where s_i is a latitude/longitude.

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

would have spatially correlated errors ϵ_i.

It is common to model this correlation as

$$Cov(\epsilon_i, \epsilon_j) = \sigma^2 exp(-d_{ij}/\phi)$$

ϕ: controls exponential decay of correlation as distance between sites d_{ij} increases,

Likelihood: $Y|\beta, \sigma^2, \rho \sim N(X\beta, \sigma^2\Sigma(\phi))$
$\Sigma(\phi)$ has (i,j) relevant $exp(-d_{ij}/\phi_{ij})$
$_{n\times n}$
Prior for ϕ: Uni(0, max(distance point))

See spatial$_{\text{modelingex.pdf}}$ for more details. It's really good.

8.5 Flexible Regression Modeling

Bayesian can't be *truly* Non-parametric because a Likelihood distribution needs to be specified in order to get a posterior, but one can make as generic a distribution as possible.

Goal: Specify a model that is so flexible that it certainly captures the *true* model.

A Bayesian Non-parametric (BNP) model may have infinite parameters.

Useful in cases where NP models are difficult conceptually and computationally so these Semi-parametric models can provide good approximations.

8.6 Turning a Parametric Model Non-paramatric

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ \epsilon \sim N(0, \sigma^2)$$

Assumptions

1. ϵ_i are independent
2. ϵ_i are Gaussian
3. Mean of Y_i is linear in X
4. The residual distribution doesn't depend on X

Alternatives

1. Parametric Alternatives such as a Time Series model.
2. Let $\epsilon_i \sim F$ and place a prior on the distribution F.
3. Let $E(Y|X) = g(X)$ and put a prior on function g.
4. Hereoskedastic regression $Var(\epsilon_i) = exp(\alpha_0 + \alpha_1 X)$

Relax assumption about Linearity in the Mean (3). Non-parametric regressions puts a prior on the curve rather than the parameters that determine the model.

8.7 Semiparametric Regression

Approximate the function using a finite basis expansion

$$g(X) = \sum_{j=1}^{J} \mathbb{B}_j(X)\beta_j$$

$\mathbb{B}_j(X)$: Known basis functions
β_j: Unknown coefficients that determine the shape of g.
Ex
Cubic Spline Basis Function: $\mathbb{B}_j(X) = (X - \nu_j)_+^3$
ν_j: fixed knots that span the range of X.
Other basis expansions include Wavelets, Fourier, etc.

A basis expansion of J terms can match the true curve f at any J points $X_1, ..., X_j$. So? Increasing J gives an arbitrary flexible model!

8.7.1 Fitting a Model

The model is $Y_i \sim N(\mathbb{B}_i^T\beta, \sigma^2)$ where $\beta_j \sim N(0, \tau^2)$ and $\mathbb{B}_i$ is comprised of known basis functions $\mathbb{B}_j(X_i)$

- How to pick J?
- Can we have more basis functions than observations?
- What would you do if your prior was g that was probably quadratic, but you are not 100% sure?

> These questions were not answered. Get them from office hours?

9 Model Comparisons & Hierarchical Linear Models (2020/11/05)

9.1 Model Comparisons

Deviance, Adjusted R^2, AIC, BIC are *frequentist* statistics for comparing models.

For a given dataset, how do we determine whether a simple model is sufficient or if we need to bring in "another tool"?

A statistical model is a mathematical representation of the system that includes and biases in the observation process.

Ideally, we want a simple and accurate model.

9.1.1 Bayesian Model Criteria

- Bayesian Factors
- Stochastic Search Variable Selection
- Cross Validation
- Deviance Information Criteria (DIC)
- WAIC

1. Bayesian Factors (Bayesian Hypothesis Test)

   Gold standard.

   Example

   M_1, M_2 are models

   $Y \sim Bin(n, \theta)$

   $M_1 : \theta = 0.5$

   $M_2 : \theta \neq 0.5$

   Compute the posterior of each model. This requires prior probabilities $P(M_1)$ and $P(M_2)$. Note that this is a prior **on the model** instead of on the parameters!.

   $$BF = \frac{PosteriorOdds}{PriorOdds} = \frac{P(M_2|Y)/P(M_1|Y)}{P(M_2)/P(M_1)} = \frac{P(Y|M_2)}{P(Y|M_1)}$$

   Rule of Thumb

   - BF > 10, strong evidence for M_2
   - BF > 100, decisive evidence for M_2

   Using the above example.

   $\theta \sim Beta(a, b)$

$$
\begin{aligned}
P(Y|M_1) =& \binom{n}{Y} 0.5^Y (1-0.5)^{n-Y} = \binom{n}{Y} 0.5^n \\
P(Y|M_2) =& \int_0^1 P(Y,\theta) d\theta = \int_0^1 P(Y|\theta) P(\theta) d\theta \\
=& \int_0^1 \left[\binom{n}{Y} \theta^Y (1-\theta)^{n-Y} \right] \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \right] d\theta \\
=& \binom{n}{Y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{Y+a-1} (1-\theta)^{n-Y+b-1} d\theta \text{ (Beta(Y + a, n - Y + b))} \\
=& \binom{n}{Y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(n+a+b)} \int_0^1 \frac{\Gamma(n+a+b)}{\Gamma(Y+a)\Gamma(n-Y+b)} \theta^{Y+n-1} (1-\theta)^{n-Y+1-}
\end{aligned}
\tag{37}
$$

So?

$P(Y|M_2) = \binom{n}{Y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(n+a+b)}$

Then:

$BF = \frac{P(Y|M_2)}{P(Y|M_1)} = \frac{\Gamma(a+b)\Gamma(Y+a)\Gamma(n-Y+b)}{\Gamma(a)\Gamma(b)\Gamma(n+a+b)\times 0.5^n}$

2. Issues with Bayes Factors

   - Can be hard to compute integrals which is only feasible for simple models.
   - Requires proper priors
   - Can be very sensitive to Priors (Lindsey's Paradox)

3. Bayes Factors - MCMC

   If models can be written as nested, then MCMC can be used to approximate model probabilities.

   $M_1 : E(Y) = \beta_0 + \beta_1 X$

   $M_2 : \beta_0 + \beta_1 X_1 + \beta_2 X^2$

   Then the model can be written as $E(Y) = \beta_0 + \beta_1 X + \gamma\beta_2 X^2$ where $\gamma \in \{0, 1\}$ and γ is the model.

   The prior on the model becomes $\gamma \sim Bern(0.5)$ which then $P(\gamma = 1|Y)$ and $P(\gamma = 0|Y)$ can be approximated using MCMC.

9.1.2 Stochastic Search Variable Selection

Bayesian equivalent of forward, backward, and stepwise selection.

Place a prior on all 2^P models us P variable inclusion indicators γ_j

Very Sensitive to priors.

9.1.3 Model Averaging

$M_1 : E(Y) = \beta_0 + \beta_1 X$

$M_2 : \beta_0 + \beta_1 X_1 + \beta_2 X^2$

Situation: We prefer to use M_1 because it is simpler but both are equally likely. If M_1 is chosen, we are suppressing uncertainty about the model form.

Let $\hat{Y}_k$ be the prediction from Model M_k for $k = 1, 2$

$$\hat{Y} = w\hat{Y}_1 + (1 - w)\hat{Y}_2$$

The optimal weight w is the posterior probability of M_1. This adds stability to the model.

Is this the same as an ensemble model?

9.1.4 Cross Validation

Same approach as in classical statistics.

Operates under the assumption that the "true" model likely produces better out-of-sample predictions than competing models.

1. K-Fold Cross Validation
   (a) Split data into K Groups
   (b) Set K aside as the test set and fit on K- 1 groups
   (c) Make prediction son test set K
   (d) Repeat 1-2 for K = 1, . . . , K
   (e) Measure Prediction Accuracy

   Variants

   - k = 5, 10
   - k = n. i.e. leave-one-out validation
   - $\hat{Y}_i$ can be the posterior predictive mean or median.
   - MSE can be replaced with MAD.

9.1.5 Information Criteria

Usually a function of the deviance: $P(Y|\theta) = -2log\left[f(Y|\theta)\right]$

Ideally a small deviance is chosen.

If a model is too complex with a small deviance, overfitting can happen.

$$AIC = D(Y|\hat{\theta}) + 2p,\ \hat{\theta} = MLE$$

$$BIC = D(Y|\hat{\theta}) + log(n)p$$

The BIC is an approximation to the log Bayes Factor of the model compared to the null model. However, it is only good for large n.

1. Deviance Information Criteria (DIC)

   - popular among Bayesians
   - Only fits one model
   - Can be applied to complex models

   But! Only applies when posterior is approximately normal. Also, tends to favor overly complex models. Can only be used to compare models with the same likelihood.

   Effective Number of Parameters is a way to measure model complexity.

   Let $\bar{D} = E[D(Y|\theta)]$ and $\hat{\theta}$ be the posterior mean of θ

   Effective number of parameters: $P_D = \bar{D} - D(Y|\hat{\theta})$

   Thus

   $$DIC = \bar{D} + P_D = D(Y|\hat{\theta}) + 2P_D$$

   Want a small $\bar{D}$ and small P_D

   $P_D \approx P$ when uninformative priors are used. $P_D << P$ with strong priors.

   Rule of Thumb

   Difference less than 5 is not definitive. Difference greater than 10 is substantial.

   (a) DIC - One Way Random Effects Model (Proof)
   $Y_{ij} = \mu_j + \epsilon_{ij}$ where

- Y_{ij}: observation i for subject j
- μ_j: mean for subject j
- $i = 1, ..., n$
- $j = 1, ..., p$

$\epsilon_{ij} \sim N(0, \tau_\epsilon)$ where $\tau_\epsilon = \frac{1}{\sigma^2}$
$\mu_j \sim iidN(0, \tau_\mu)$ where $\tau_\mu = \frac{1}{\sigma_\mu^2}$
We assume τ_ϵ, τ_μ are known

Recall: $\mu_j|\vec{Y} \sim N(E_j, P_j)$ where $E_j = \frac{n\tau_\epsilon}{n\tau_\epsilon + \tau_\mu}\bar{Y}_j$ and $P_j = n\tau_\epsilon + \tau_\mu$

Deviance: $D(\vec{Y}|\vec{\mu}) = -2log[f(\vec{Y}|\vec{\mu})]$
So?

$$\begin{aligned} f(\vec{Y}|\vec{\mu}) =& \prod_{i=1}^{n}\prod_{j=1}^{p} \frac{\tau_\epsilon^{0.5}}{\sqrt{2\pi}} exp\left[-\frac{\tau_\epsilon}{2}(Y_{ij} - \mu_j)^2\right] \\ =& \left[\frac{\tau_\epsilon}{\sqrt{2\pi}}\right]^{np/2} exp\left[-\frac{\tau_\epsilon}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - \mu_j)^2\right] \qquad (38) \\ logf(\vec{Y}|\vec{\mu}) =& \frac{np}{2} log\left[\frac{\tau_\epsilon}{\sqrt{2\pi}}\right] - \frac{\tau_\epsilon}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - \mu_j)^2 \end{aligned}$$

So?

$$\begin{aligned} D(\vec{Y}|\vec{\mu}) =& -nplog\left[\frac{\tau_a}{2\pi}\right] + \tau_\epsilon\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - \mu_j)^2 \\ =& -nplog\left[\frac{\tau_a}{2\pi}\right] + \tau_\epsilon\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - E_j)^2 \\ \bar{D} = E_{\hat{\mu}|\hat{Y}}[D(\vec{Y}|\vec{\mu})] =& -nplog\left[\frac{\tau_a}{2\pi}\right] + \tau_\epsilon E\left[\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - \mu_j)^2\right] \end{aligned} \qquad (39)$$

Note: $E(\mu_j) = E_j E(\mu_j^2) = E(\mu_j^2) + Var(\mu_j) = E_j^2 + \frac{1}{P_j}$

$$
\begin{aligned}
P_D = \bar{D} - D(\vec{Y}|\vec{\mu}) =& \tau_\epsilon E\left[\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij}-\mu_j)^2\right] - \tau_\epsilon \sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij}-E_j)^2 \\
=& \tau_\epsilon \left[E\left[\sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij}-\mu_j)^2\right] - \sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij}^2 - 2Y_{ij}E_j + E_j^2)\right] \\
=& \tau_\epsilon \left[\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}^2 - \sum_{i=1}^{n}\sum_{j=1}^{p}2Y_{ij}\mu_j + \sum_{i=1}^{n}\sum_{j=1}^{p}E(\mu_j^2) - \sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}^2 + \sum_{i=1}^{n}\sum_{j=1}^{p}2\right. \\
=& \tau_\epsilon \left[\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}^2 - \sum_{i=1}^{n}\sum_{j=1}^{p}2Y_{ij}E_j + \sum_{i=1}^{n}\sum_{j=1}^{p}(E_j + \frac{1}{P-j}) - \sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}^2 + \right. \\
=& \tau_\epsilon \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{1}{P_j} \\
=& \tau_\epsilon \sum_{j=1}^{p}\frac{n}{P_j} \\
=& \tau_\epsilon n \sum_{j=1}^{p}\frac{1}{n\tau_\epsilon + \tau_\mu} \\
\rightarrow& P\frac{n\tau_\epsilon}{n\tau_\epsilon + \tau_\mu}
\end{aligned}
\tag{40}
$$

What this proves

i. $0 < P_d < P \forall n, \tau_\epsilon, \tau_\mu$
ii. $\tau_\mu = 0$ gives a flat prior. $P_D = P$
iii. $\tau_\mu = \infty$ gives a tight prior. $P_D = 0$

9.1.6 Watanabe-Akaike Information Criteria

Alternative to DIC

Gives similar results but is arguably more theoretically justified.

Aims to be an approximation of leave-one-out CV.

Written in terms of the posterior of the likelihood rather than the parameters.

Let m_i and v_i
$)betheposteriormean, varianceoflog[f(Y_i|\theta)]$

Effective Model Size: $p_W = \sum_{i=1}^{n} \nu_i$

$$WAIC = -2\sum_{i=1}^{n} m_i + 2p_W$$

9.1.7 Posterior Predictive Checks

9.1.8 Posterior Predictive Distribution

1. After comparing a few models, choose one that fits best.
2. Verify model is adequate.
   - QQ Plots, added variable plts, etc.
3. Posterior Predictive Check. Leads to a Bayesian p-value.

Need Posterior Predictive Distribution

$$f(Y_{new}|Y) = \int f(Y_{new} m\theta|Y)d\theta = \int f(Y_{new}|\theta)f(\theta|Y)d\theta$$

To make S draws from the PPD, for each S of the MCMC draws of θ, we draw Y_{new}. This gives draws from the PPD and clearly accounts for uncertainty in θ

9.1.9 Posterior Predictive Checks

Sample many datasets from PPD with identical design (Same X. Same n).
Define a statistic describing the Dataset: $d(Y) = max\{Y_!, ..., Y_n\}$
Statistic from original dataset: d_0
Statistic from simulated dataset s: d_s
If the model is correct, d_0 falls between d_1 and d_s

1. Bayesian P-value

   A measure of how extreme the observed data are relative to this sampling distribution

$$p = \frac{1}{\mathbb{S}} \sum_{s=1}^{\mathbb{S}} I(d_s > d_0)$$

   If p is near 0 or 1, the model does not fit.

   This is repeated for several d to give a comprehension for model fit.

10 Hierarchical Models & Closing Thoughts (2020/11/12)

10.1 Hierarchical Models

provides a framework for building complex and high-dimensional models from simple and low-dimensional building blocks.

Works well since MCMC is conducive to Hierarchical Models.

Items to cover:

1. Building Hierarchical Models through Layers.
2. DAGS

10.1.1 Layers

1. Data Layer: $[Y|\theta, \alpha]$ is the likelihood for observed data Y given the model parameters.
2. Process Layer: $[\theta, \alpha]$ model for the parameters θ that define the latest data generating process.
3. Prior: $[\alpha]$ prior for the hyper-parameters.

1. Example

   Data Layer

   Let

   - S_t: number susceptible at time t
   - I_t: Infected individuals at time t
   - Y_T: number of observed cases at time t

   Data Layer models our ability to measure I_t. $Y_T|I_T \sim Bin(I_t, \rho)$.

   Assume no false positives and false negative probability, ρ.

   What do I need to know to figure out I_t?

   Process Layer

   Reed-Frost Model

   $I_{t+1} \sim Bin(S_t, 1 - (1-q)^{I_t})$

   $S_{t+1} = S_t - I_{t+1}$

   Prior Layer

$$\begin{aligned} I_1 &\sim Poisson(\lambda) \\ S_1 &\sim Poisson(\lambda_2) \\ \rho, \alpha &\sim Beta(a, b) \end{aligned} \tag{41}$$

When to stop adding models? It can go down a dangerous rabbit hole.

Rule of Thumb: Be careful assigning priors to parameters in layers without replication (aka multiple observations).

10.1.2 DAGS

Graphical representation of a hierarchical model. Sometimes referred to as a Bayesian Network.

Each observation and parameter is a node.

$X \to Y$: Y depends on X.

1. Misc Building a model this way always has a valid joint distribution. ex: $f(x, y, z) = f(x)f(y|x)f(z|x, y)$ is a valid DAG.

   Adhoc constructions may not give valid distributions.

   How does Data from Y_3 affect Y_2? It doesn't because it's a DAG.

   MCMC is efficient even when the number of parameters is large. This is because only connected nodes need to be updated.

10.1.3 Random Slopes Model

$Y_{ij} \sim N(\gamma_{0i} + \gamma_{1i}X_i, \sigma^2)$ where i: subject and j: observation.

$Y_i \sim N(\vec{\beta}, \Sigma)$

Aside: What are good prior hyperparameters for an Inverse Wishart? It is a subject of much debate. It is thought that choosing R to be an identity matrix is a good idea. Or choosing K = p + 1 which leads to infinitesimal spread. There isn't generally a **good** rule of thumb.

10.1.4 Missing Data Models

$Y_i \sim N(\beta_0 + \Sigma\beta_i X_i, \sigma^2)$

1. Missing Responses

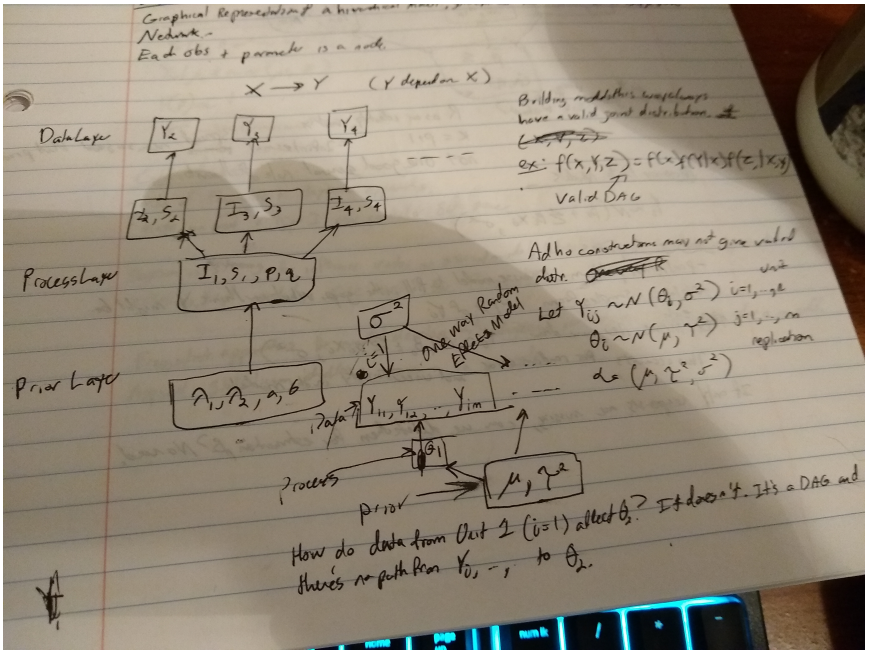

Figure 1: Bayesian Model DAG

- Prediction Problem. Can use model to fill in the gaps. What do you think Y might be?
- Obtain samples from PPD of Y_i
- for each MCMC iteration, we draw $Y_i \sim N(\beta_0 + \Sigma\beta_i X_i, \sigma^2)$. This accounts for random error and uncertainty in the model.
- If only responses are missing, can we delete them for estimation? No need unless there are patterns in the responses that might be missing. For example, missing temperature data when trying to ascertain whether a heatwave has occurred.

2. Missing Covariates

- Simplest approach is imputation, but doesn't account for uncertainty.
- MCMC can handle this as well.

The main idea is to treat the missing values as unknown parameters. Unknown parameters need priors so missing $X_i = (X_{i1}, ..., X_{ip})^T$ must have priors $X_i \sim N(...)$

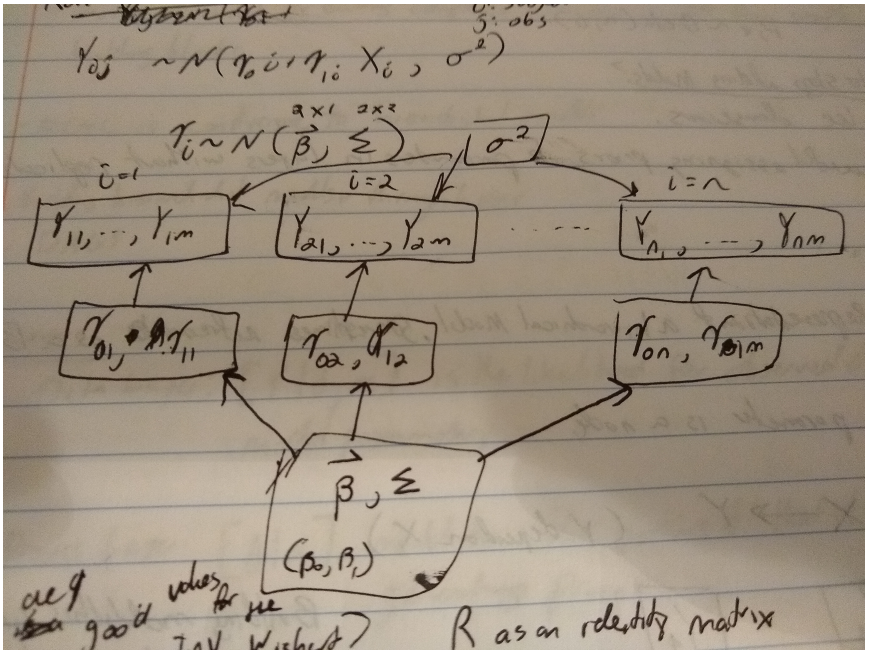

Figure 2: Random Slopes Model DAG

If priors are way off, then bad results will be returned. If specified correctly, lead to inference of β's in the model.

Assumptions about Missing Data

- Missing Status is independent of Y and X
- Covariates are Gaussian

There are ways to relax both assumptions but it becomes complicated.

> If Data are not Missing At Random (MAR), the Bayesian Model will likely give bad results.

3. Hierarchical LR Model with Missing Data

   $Y_i|X_i, \beta, \sigma^2 \sim N(X_i^T\beta, \sigma^2)$

   $X_i|\mu, \Sigma \sim N(\mu, \Sigma)$

   $P(\beta) \propto 1$

   $\sigma^2 \sim InvGamma(0.01, 0.01)$

10.1.5 Closing Thoughts

1. Calibrated Bayes

   - Mix of Frequentist and Bayesian Approaches
   - Maybe we want a Type 1 error control or 80% Power.

   Bayesian is strong for influence under an assumed model, but relatively weak for development and assessment of models. Vice versa for frequentist approaches.

   Frequentist approaches for model development and Bayesian methods for inference under a model are ideal in principle. Applied statisticians should be Bayesian in principle.

2. Bayesian Decision Theory

   - Need to determine the "best" Bayes method.

   Should we take posterior mode, median, or mean?

   Need a scoring system.

   Let $l(\hat{\theta}, \theta)$ be a loss function.

   $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$: Squared Loss function (L2 norm?)

   $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$: Absolute Loss function (L1 Norm?)

   Summary of the posterior that minimizes the expected (posterior) loss is Bayes Rule. However, Hypothesis testing requires a more complicated loss function.

3. Bayes Rule for Squared-error loss

   Loss: $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$

   $\bar{\theta} = E(\theta|Y)$: Posterior Mean of θ

   Expected Loss

$$\begin{aligned} E_{\theta|Y}[l(\hat{\theta}, \theta)] =& E[(\theta\hat{-}\theta)^2] \\ =& E[(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)] \\ =& E[(\theta - \bar{\theta})^2 + 2(\hat{\theta} - \bar{\theta})(\bar{\theta} - \hat{\theta}) + (\bar{\theta} - \hat{\theta})^2] \\ =& E[(\theta - \bar{\theta})^2] + 2(\bar{\theta} - \hat{\theta})E[(\hat{\theta} - \bar{\theta})] + (\bar{\theta} - \hat{\theta})^2 \\ =& Var(\theta|Y) + (\bar{\theta} - \hat{\theta})^2 \end{aligned} \tag{42}$$

What value of $\hat{\theta}$ minimizes $E[l(\theta, \hat{\theta})]$?

$\therefore \hat{\theta} = \bar{\theta}$ minimizes the expected loss, so posterior mean is Bayes Rules

4. Bayes Rule for Hypothesis Testing

   Let $\theta = 0$ if H_0 is true. Fail to reject the null hypothesis Let $\theta = 1$ if H_A is true. Reject the null hypothesis.

   P_0: Posterior probability of H_0. $P(\theta = 0|Y)$.

   $1 - P_0$: Posterior Probability of H_A. $P(\theta = 1|Y)$

$$l(\theta, \hat{\theta}) = \begin{cases} \lambda_1, & \text{Type 1 Error}, \hat{\theta} = 1, \theta = 0 \\ \lambda_2, & \text{Type 2 Error}, \hat{\theta} = 0, \theta = 1 \end{cases}$$

   If $\hat{\theta} = 1$, expected loss function is $P_0\lambda_1$ If $\hat{\theta} = 0$, expected loss function is $(1 - P_0)\lambda_2$

$$P_0\lambda_1 < (1 - P_0)\lambda_2 \equiv P_0 < \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

5. Bias Variance Tradeoff

   $Y_i \sim N(\mu, \sigma^2)$

   Estimator 1: $\hat{\mu}_1 = \bar{Y}$

   Estimator 2: $\hat{\mu}_2 = c\bar{Y}$ where $c = \frac{n}{n+m}$

   $\hat{\mu}_2$ is the posterior mean under prior $\mu \sim N(0, \frac{\sigma^2}{m})$.

   Compute the Bias, Variance, MSE of each estimator (MSE = bias^2 + Var).

   Which is preferred?

   Bias

$$\begin{aligned} E(\mu_1 - \mu) =& E(\bar{Y}) - \mu \\ =& E(\frac{1}{n}\sum_1^n Y_i) - \mu \\ =& \frac{1}{n}E(\sum_1^n Y_i) - \mu \\ =& \frac{1}{n}n\mu - \mu = 0 \end{aligned} \tag{43}$$

$$E(\mu_2 - \mu) = cE(\bar{Y}) - \mu = c\mu - \mu = \mu(c-1) \tag{44}$$

As m increases, c becomes less than 1 and will shrink the variance, but more bias will be introduced.

> In a Bayesian Hypothesis test, λ_1, λ_2 can be set to control Type 1 error.

Bayesian estimators have smaller SE because the prior adds information to the model. Bayesian estimators are biased if prior is not centered on the truth. With a weak prior, the outcomes between a Frequentist approach and a Bayesian Approach are similar.

6. Bayesian CLT

Assumption

- Usual MLE conditions on the likelihood.
- Prior doesn't depend on n and puts non-zero prob on the true value θ_0.

Then

$$P(\theta|Y) \to N(\theta_0, I(\theta_0)^{-1})$$

Bayes methods asymptotically unbiased.

Bayes and MLE equivalent with Large Samples but interpretations are different.

Bayes CLT useful for initial values and tuning.

For each regression coefficient, you should have 10x the number of records. When the model is sparse, BLR has much smaller MSE than OLS.