Bayesian Data Analysis

Module 3: Models with more than one parameter

Stat 474/574

Motivation

- Most realistic problems are better summarized by models with more than one parameter.
- Typically, we are interested in making inference about one or a few of those parameters.
- The frequentist approach to inference in multi-parameter models typically consists in:
    1. Maximizing a joint likelihood, which can get difficult when there are many model parameters, or
    2. Proceeding in steps: first estimate some of the parameters and then plug those estimates in the model to estimate the rest.
- In the Bayesian approach, inferences are based on the *marginal posterior distributions* of the parameters of interest.
- Parameters that are not of interest are called *nuisance parameters*.

Nuisance parameters

- Consider a model with two parameters (θ_1, θ_2) (e.g., a normal distribution with unknown mean and variance).
- Suppose that we are interested in θ_1 so that θ_2 is a nuisance parameter.
- The marginal posterior distribution of interest is $p(\theta_1|y)$, which can be obtained directly from the *joint posterior density*:

$$p(\theta_1, \theta_2|y) \propto p(\theta_1, \theta_2)p(y|\theta_1, \theta_2)$$

by integrating with respect to θ_2:

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2.$$

Nuisance parameters (cont'd)

- Note too that

$$p(\theta_1|y) = \int p(\theta_1, |\theta_2, y)p(\theta_2|y)d\theta_2.$$

- The marginal of θ_1 is a **mixture of conditionals** on θ_2, or a **weighted average** of the conditional distribution of θ_1 evaluated at different values of θ_2. Weights are given by the marginal distribution $p(\theta_2|y)$.

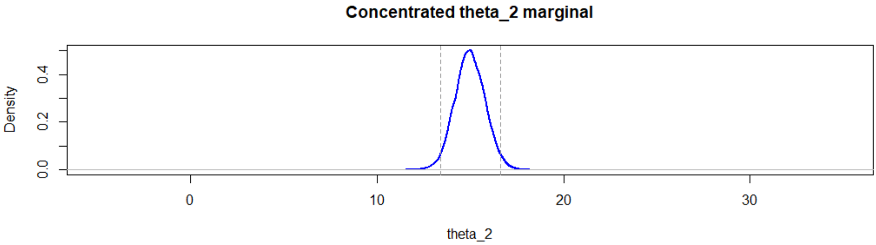
Nuisance parameters (cont'd)

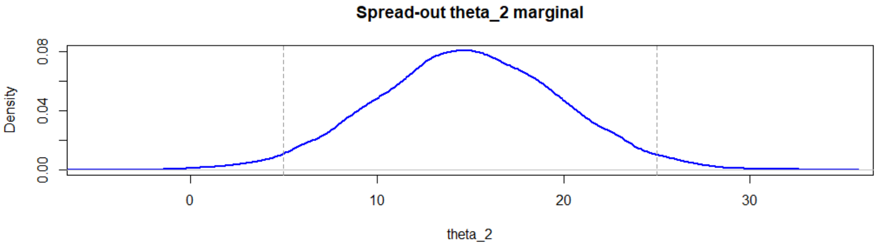
- This is a critical difference between Bayesians and frequentists.
- By averaging the conditional $p(\theta_1, |\theta_2, y)$ over possible values of θ_2, we explicitly recognize our uncertainty about the true value of θ_2.
- For illustration, consider two extreme cases:
    1. We are almost certain about the true value of θ_2: If the prior and the sample are very informative about θ_2, then the marginal $p(\theta_2|y)$ is concentrated around some value $\hat{\theta}_2$. In that case,

    $$p(\theta_1|y) \approx p(\theta_1|\hat{\theta_2}, y).$$

    2. We do not have much information about θ_2. In this case the marginal $p(\theta_2|y)$ will assign relatively high probability to a wide range of values of θ_2. Point estimate $\hat{\theta}_2$ is no longer meaningful and therefore it is important to average $p(\theta_1|y, \theta_2)$ over the range of likely values of θ_2.

To visualize...





Nuisance parameters (cont'd)

- In most cases, the integral is not computed analytically.
- Instead, we use a two-step simulation approach:
  1. Marginal simulation step: Draw value $\theta_2^{(k)}$ of θ_2 from $p(\theta_2|y)$ for $k = 1, 2, ...$
  2. Conditional simulation step: For each $\theta_2^{(k)}$, draw a value of θ_1 from the conditional density $p(\theta_1|\theta_2^{(k)}, y)$.
- This is an effective approach when both the marginal and conditional distributions are of standard form.
- In general, we will need more sophisticated simulation approaches, which we will learn later.

Example: Normal model

- Let y_i *iid* from $N(\mu, \sigma^2)$, both unknown.
- Suppose that we choose a non-informative prior for (μ, σ^2) and assume prior independence of both parameters so that:

$$p(\mu, \sigma^2) \propto 1 \times \sigma^{-2}.$$

- The joint posterior distribution is:

$$\begin{aligned} p(\mu, \sigma^2 | y) &\propto p(\mu, \sigma^2) p(y | \mu, \sigma^2) \\ &\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right). \end{aligned}$$

Example: Normal model (cont'd)

- Note that

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\mu)^2 &= \sum_i (y_i^2 - 2\mu y_i + \mu^2) \\
&= \sum_i y_i^2 - 2\mu n\bar{y} + n\mu^2 \\
&= \sum_i (y_i-\bar{y})^2 + n(\bar{y}-\mu)^2,
\end{aligned}
$$

by adding and subtracting $2n\bar{y}^2$.

Example: Normal model (cont'd)

- Let
$$s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2.$$
- Then we can write the joint posterior for (μ, σ^2) as
$$p(\mu, \sigma^2|y) \propto \sigma^{-n-2}\exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]).$$
- Little aside: the sufficient statistics for μ, σ^2 are $(\bar{y}, s^2)$,

Example: Conditional posterior $p(\mu|\sigma^2, y)$

- Conditional on σ^2:

$$p(\mu|\sigma^2, y) = N(\bar{y}, \sigma^2/n).$$

- We know this from the earlier chapter (posterior of normal mean when variance is known).
- We can also see this by noting that, viewed as a function of μ only:

$$p(\mu|\sigma^2, y) \propto \exp(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2),$$

which we recognize as the kernel of a $N(\bar{y}, \sigma^2/n)$.

Marginal posterior $p(\sigma^2|y)$

- To get $p(\sigma^2|y)$ we need to integrate $p(\mu, \sigma^2|y)$ over μ:

$$
\begin{aligned}
p(\sigma^2|y) &\propto \int \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]) d\mu \\
&\propto \sigma^{-n-2} \exp(-\frac{(n-1)s^2}{2\sigma^2}) \int \exp(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2) d\mu \\
&\propto \sigma^{-n-2} \exp(-\frac{(n-1)s^2}{2\sigma^2}) \sqrt{2\pi\sigma^2/n}.
\end{aligned}
$$

- Then

$$p(\sigma^2|y) \propto (\sigma^2)^{-(n+1)/2} \exp(-\frac{(n-1)s^2}{2\sigma^2}),$$

  which is proportional to a *scaled-inverse* χ^2 distribution with degrees of freedom $(n-1)$ and scale s^2.

- Do you recall the classical result? Conditional on σ^2, the distribution of (the scaled sufficient statistic) $(n-1)s^2/\sigma^2$ is χ^2_{n-1}.

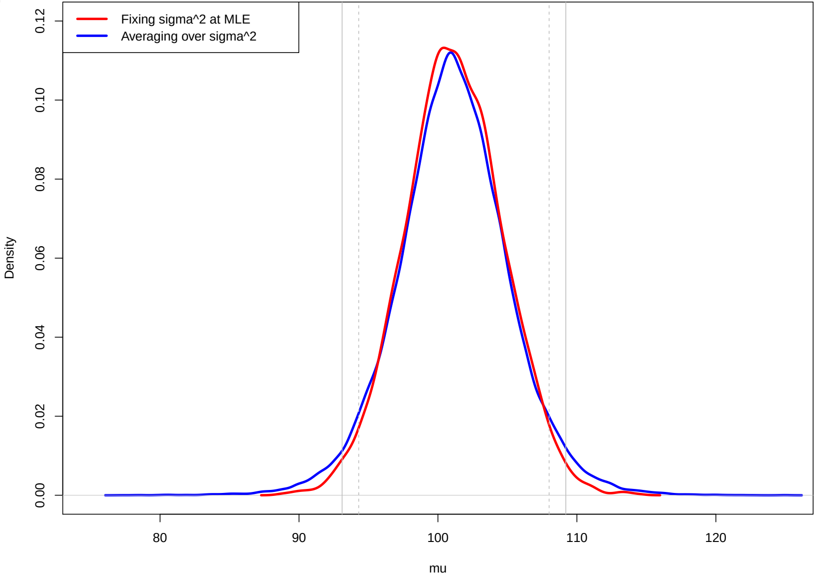
Toy example

- $y \sim \mathrm{N}(\mu, \sigma^2)$ and the joint prior is the simple non-informative $p(\mu, \sigma^2) \propto \sigma^{-2}$.
- We draw values of μ from $p(\mu|\sigma^2, y)$ using two different methods:
  - Act as a frequentist: first get $\hat{\sigma}^2$, the MLE of σ^2 and then draw μ from

$$p(\mu|\sigma^2, y) \approx p(\mu|\sigma^2 = \hat{\sigma}^2, y).$$

  - Act as a Bayesian and integrate σ^2 out of the joint posterior distribution $p(\mu, \sigma^2|y)$
    1. Draw $\sigma_1^2, ..., \sigma_M^2$ from $p(\sigma^2|y)$.
    2. Draw M μs from $p(\mu|\sigma_k^2, y)$, $k = 1, ..., M$.

Estimated conditional posteriors



Aside: Analytical derivation in the normal model

- For the normal model, we can derive the marginal $p(\mu|y)$ analytically:

$$
\begin{aligned}
p(\mu|y) &= \int p(\mu, \sigma^2|y) d\sigma^2 \\
&\propto \int (\frac{1}{2\sigma^2})^{n/2+1} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]) d\sigma^2.
\end{aligned}
$$

- Use the transformation

$$
z = \frac{A}{2\sigma^2}
$$

where $A = (n-1)s^2 + n(\bar{y}-\mu)^2$.

Aside (cont'd)

- Then
$$\frac{d\sigma^2}{dz} = -\frac{A}{2z^2}$$
and
$$\begin{aligned} p(\mu|y) &\propto \int_0^\infty (\frac{z}{A})^{\frac{n}{2}+1} \frac{A}{z^2} \exp(-z) dz \\ &\propto A^{-n/2} \int z^{\frac{n}{2}-1} \exp(-z) dz. \end{aligned}$$
- Integrand is unnormalized Gamma(n/2, 1), so integral is constant w.r.t. μ
- Recall that $A = (n-1)s^2 + n(\bar{y}-\mu)^2$. Then
$$\begin{aligned} p(\mu|y) &\propto A^{-n/2} \\ &\propto [(n-1)s^2 + n(\bar{y}-\mu)^2]^{-n/2} \\ &\propto [1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}]^{-n/2} \end{aligned}$$

Aside (cont'd)

- For the non-informative prior $p(\mu, \sigma^2) \propto \sigma^{-2}$, the posterior distribution of μ is a non-standard t. Then,

$$p(\frac{\mu - \bar{y}}{s/\sqrt{n}}|y) = t_{n-1}$$

  the standard t distribution.
- Notice similarity to classical result: for iid normal observations from $N(\mu, \sigma^2)$, given (μ, σ^2), the *pivotal quantity*

$$\frac{\bar{y} - \mu}{s/\sqrt{n}}|\mu, \sigma^2 \sim t_{n-1}$$

- A *pivot* is a non-trivial function of the data and the parameter(s) θ whose distribution, given θ, is independent of θ. Property deduced from *sampling distribution* as above.

Posterior predictive of future observations

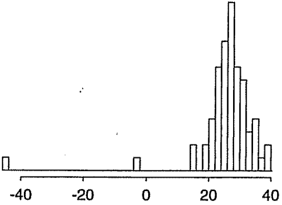
- The posterior predictive distribution of a future observation $\tilde{y}$ is a *mixture*:

$$p(\tilde{y}|y) = \int\int p(\tilde{y}|y, \sigma^2, \mu) p(\mu, \sigma^2|y) d\mu d\sigma^2.$$

- The first factor in the integrand is just the normal model, and it does not depend on y at all.
- To simulate $\tilde{y}$ from its posterior predictive distributions, do the following:
  1. Draw σ^2 from Inv-$\chi^2(n-1, s^2)$,
  2. Then draw μ from $N(\bar{y}, \sigma^2/n)$, and finally
  3. Draw $\tilde{y}$ from $N(\mu, \sigma^2)$.

Example: Estimating the speed of light

- In 1882, Simon Newcomb set up an experiment to measure the speed with which light travels a distance of 7,442 meters. He took 66 measurements and expressed them as a deviation from some constant.



Speed of light (cont'd)

- Assume (inappropriately?) that the 66 measurements are iid $N(\mu, \sigma^2)$. We are interested in making inferences about μ, the mean speed of light.
- The sample mean $\bar{y}$ is 26.2 and the sample standard deviation s is 10.8.
- If we assume a non-informative prior where

$$p(\mu, \sigma^2) \propto \sigma^{-2},$$

we get a 95% credible set equal to $[23.6,\ \ 28.8]$ for μ.

Speed of light (cont'd)

- To get the credible set for μ, we proceeded by simulation:
  1. First draw σ^2 from the Invχ^2 distribution with 65 degrees of freedom. To do this, we first draw a random variable from a χ^2 with 65 degrees of freedom and then get $\sigma^2 = 65s^2/\text{draw}$.
  2. **Alternative:** Load the package `geoR` which includes the function `rinvchisq` to draw directly from the Invχ^2.
  3. We then plug this value into a normal distribution with mean 26.2 and variance $\sigma^2/66$ and draw a value of μ.
- Based on today's technology, the speed of light (in the same scale as Newcomb's measurements) is 33.0, which falls outside of the 95% credible set obtained from Newcomb's measurements.
- Problem is the initial sampling model for the measurements, which is not really correct.

Conjugate prior for the normal model

- Recall that using a non-informative prior, we found that

$$
\begin{aligned}
p(\mu|\sigma^2, y) &\propto N(\bar{y}, \sigma^2/n) \\
p(\sigma^2|y) &\propto Inv-\chi^2(n-1, s^2).
\end{aligned}
$$

- Then, factoring $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$ the conjugate prior for σ^2 would also be scaled inverse χ^2 and for μ (conditional on σ^2) would be normal.

Conjugate prior for the normal model (cont'd)

- Consider

$$
\begin{aligned}
\mu|\sigma^2 &\sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0) \\
\sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2).
\end{aligned}
$$

- Jointly:

$$
p(\mu, \sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2]).
$$

Conjugate prior for the normal model (cont'd)

- Note that μ and σ^2 are not independent a priori in the joint conjugate prior.
- The posterior density for (μ, σ^2) is obtained as follows:
  - Multiply the sampling distribution by the N-Inv-$\chi^2(\mu_0, \sigma^2/\kappa_0; \nu_0, \sigma_0^2)$ prior.
  - Expand the two squares in μ.
  - Complete the square by adding and subtracting a term that depends on $\bar{y}$ and μ_0.

Conjugate prior for the normal model (cont'd)

- Then, $p(\mu, \sigma^2|y) \propto$ N-Inv-$\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2)$, where

$$
\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2.
\end{aligned}
$$

Conjugate prior for the normal model (cont'd)

- Interpretation of posterior parameters:
  - As before, μ_n is a weighted average of the prior mean and the sample mean.
  - The posterior "guess" $\nu_n \sigma_n^2$ is the sum of the sample sum of squared deviations, the prior sum of squared deviations, and additional uncertainty due to the difference between the sample mean and the prior mean.

Conjugate prior for the normal model(cont'd)

- Conditional posterior of μ: As before

$$\mu|\sigma^2, y \sim \mathrm{N}(\mu_n, \sigma^2/\kappa_n).$$

- Marginal posterior of σ^2: As before

$$\sigma^2|y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2).$$

- Marginal posterior of μ: As before

$$\mu|y \sim t_{\nu_n}(\mu|\mu_n, \sigma_n^2/\kappa_n).$$

- Two ways to sample from the joint posterior distribution:
  1. Sample μ from t and σ^2 from Inv-χ^2.
  2. Sample σ^2 from Inv-χ^2 and, given σ^2, sample μ from N.

Semi-conjugate prior for normal model

- We might be inclined to set independent priors for μ and σ^2, where

$$\begin{aligned} \mu &\sim \mathrm{N}(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2). \end{aligned}$$

- This prior is *not conjugate* for the normal model and does not lead to a posterior of standard form.

Semi-conjugate prior for the normal model (cont'd)

- We can factor the joint posterior as we did earlier:

$$\mu|\sigma^2, y \sim \text{ N}(\mu_n, \tau_n^2),$$

with

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \ \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}.$$

- NOTE: Even though μ and σ^2 are independent a priori, they are not independent in the posterior.

Semi-conjugate prior and $p(\sigma^2|y)$

- The marginal posterior $p(\sigma^2|y)$ can be obtained by integrating the joint $p(\mu, \sigma^2|y)$ w.r.t. μ:

$$p(\sigma^2|y) \propto \int \mathrm{N}(\mu|\mu_0, \tau_0^2)\ \mathrm{Inv}\chi^2(\sigma^2|\nu_0, \sigma_0^2) \prod \mathrm{N}(y_i|\mu, \sigma^2) d\mu.$$

- Keeping track of normalizing constants that depend on σ^2 is messy. It is easier to note that:

$$p(\sigma^2|y) = \frac{p(\mu, \sigma^2|y)}{p(\mu|\sigma^2, y)},$$

so that

$$p(\sigma^2|y) \propto \frac{\mathrm{N}(\mu|\mu_0, \tau_0^2)\ \mathrm{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2) \prod \mathrm{N}(y_i|\mu, \sigma^2)}{\mathrm{N}(\mu|\mu_n, \tau_n^2)},$$

which is still a mess.

Semi-conjugate prior and $p(\sigma^2|y)$

- The expression in the previous page was:

$$p(\sigma^2|y) \propto \frac{\mathrm{N}(\mu|\mu_0, \tau_0^2)\ \mathrm{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2) \prod \mathrm{N}(y_i|\mu, \sigma^2)}{\mathrm{N}(\mu|\mu_n, \tau_n^2)}.$$

- If this is the marginal of σ^2, the factors that depend on μ must cancel, and therefore we know that $p(\sigma^2|y)$ does not depend on μ in the sense that we can evaluate $p(\sigma^2|y)$ for a grid of values of σ^2 and *any arbitrary* value of μ.
- Choose $\mu = \mu_n$ and then the denominator simplifies to something that is proportional to τ_n^{-1}. Then

$$p(\sigma^2|y) \propto \tau_n\ \mathrm{N}(\mu|\mu_0, \tau_0^2)\ \mathrm{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2) \prod \mathrm{N}(y_i|\mu, \sigma^2),$$

which can be evaluated for a grid of values of σ^2.

For calculations...

- Again, the expression from the previous slide is:

$$p(\sigma^2|y) \propto \tau_n \ \mathrm{N}(\mu|\mu_0, \tau_0^2) \ \mathrm{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2) \prod \ \mathrm{N}(y_i|\mu, \sigma^2),$$

which can be simplified to:

$$p(\sigma^2|y) \propto \tau_n \ \mathrm{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2) \prod \ \mathrm{N}(y_i|\mu, \sigma^2),$$

because the prior for μ does not depend on σ^2.

- We can write out the expression for $p(\sigma^2|y)$ as

$$p(\sigma^2|y) \propto \tau_n (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu_n)^2\right).$$

For calculations...(cont'd)

- By gathering similar terms, we get:

$$p(\sigma^2|y) \propto \tau_n(\sigma^2)^{-(\frac{\nu_0}{2}+\frac{n}{2}+1)} \exp\left[-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \sum_i (y_i - \mu_n)^2\right)\right].$$

- To evaluate $p(\sigma^2|y)$ on a grid of values of σ^2, remember that τ_n, μ_n also depend on σ^2.