Nuisance Parameters, Noninformative Prior Distributions, Normal Model

Frank Wood

November 16, 2009

Normal Model

Noninformative Prior Distributions

Bayesian estimation of normal mean with known variance

In many statistical analyses, estimating the mean of a normal distribution is of great importance. Here we consider estimating the mean of a normal distribution with known variance given a set of samples.

We know the likelihood of n iid observations $y = (y_1, \ldots, y_n)$, $y_i|\theta, \sigma^2 \sim N(\theta, \sigma^2)$.

$$P(y|\theta, \sigma^2) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2}$$

Inference objectives

Our Bayesian inference objectives will be, as usual, estimation of the posterior distribution over the mean

$$P(\theta|y,\sigma^2) \propto P(y|\theta,\sigma^2)P(\theta|\sigma^2)$$

and the posterior predictive distribution

$$P(y_{new}|y,\sigma^2) = \int P(y_{new}|\theta,\sigma^2)P(\theta|y,\sigma^2)d\theta$$

Gaussian distribution manipulation(s)

Because manipulations of Normal distributions is so central to statistics (and regression analysis) we highlight an important and often used technique called completing the square.

Often we will see distribution that are proportional to something like

$$p(x) \propto e^{-\frac{1}{2}(ax^2+bx+const)}$$

for some *const* that is not a function of x

When we find this we can use the method of completing the squares to identify the corresponding Normal distribution parameterization. Consider the exponent of a normal distribution

$$-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)$$

Gaussian distribution manipulation(s)

$$-\frac{1}{2}(x-\mu)\frac{1}{\tau^2}(x-\mu)$$

If we expand this out we obtain $-\frac{1}{2}(\frac{x^2}{\tau^2} - \frac{2\mu x}{\tau^2} + const)$

We can then immediately identify the mean and variance parameters of the normalized (Normal) distribution that corresponds to the original form $-\frac{1}{2}(ax^2 + bx + const)$ by inspection.

i.e. $\tau^2 = \frac{1}{a}$ and $\mu = -b\tau^2/2$.

Bayesian estimation of normal mean with known variance

Returning to the problem of estimating the mean of a normal distribution given a population of observations and a known variance we have a likelihood of

$$P(y|\theta, \sigma^2) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2}$$

and choose a prior which is normal as well, i.e.

$$P(\theta|\mu_0, \tau_0^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2}$$

this leaves us with a posterior distribution that looks like

$$P(\theta|y, \sigma^2) \propto \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2} \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2}$$

Bayesian estimation of normal mean with known variance

The expression

$$P(\theta|y,\sigma^2) \propto \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2} \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2}$$

can be simplified by throwing out the terms that don't depend on θ and moving the product into the exponent

$$P(\theta|y,\sigma^2) \propto e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\theta)^2} e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2}$$

expanding out the square, combining the exponents, and distributing the sum yields

$$P(\theta|y,\sigma^2) \propto e^{-\frac{1}{2\sigma^2}(\sum_i y_i^2 - 2\sum_i y_i\theta + n\theta^2) - \frac{1}{2\tau_0^2}(\theta^2 - 2\theta\mu_0 + \mu_0^2)}$$

Bayesian estimation of normal mean with known variance

A little more algebra yields

$$P(\theta|y,\sigma^2) \propto e^{-\frac{1}{2}\left(\frac{(\sigma^2+\tau_0^2 n)}{\sigma^2\tau_0^2}\theta^2 - 2\frac{(\sigma^2\mu_0-\tau_0^2 n\bar{y})}{\sigma^2\tau_0^2}\theta + const\right)}$$

but we can use our results from completing the square earlier ($\tau_n^2 = \frac{1}{a}$ and $\mu_n = -b\tau_n^2/2$) with

$$a = \frac{(\sigma^2+\tau_0^2 n)}{\sigma^2\tau_0^2} \text{ and } b = -2\frac{(\sigma^2\mu_0-\tau_0^2 n\bar{y})}{\sigma^2\tau_0^2}$$

to arrive at the fact that

$$\theta|y,\sigma^2 \sim N(\mu_n, \tau_n^2)$$

where

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

Intuition

If we look at the expressions for the mean and the precision for the posterior distribution

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

note that in the limit $n \to \infty$ (the number of observations grows large) the posterior mean tends towards the sample mean $\bar{y}$ and the posterior variance goes to zero.

Run

Applied Regression/normal_posterior_update/main.m

Intuition

Another way of looking at this update is in the limit as $\tau_0 \to \infty$. In this case

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

simplifies to

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} \text{ and } \mu_n = \bar{y}$$

meaning that $P(\theta|y, \sigma^2) \approx N(\theta|\bar{y}, \sigma^2/n)$. This is like assuming that $P(\theta) = const, \theta \in (-\infty, \infty)$ which is not strictly possible.

Also note that posterior distribution of θ under this improper prior is equal to the sampling distribution of the usual estimator for the mean $\hat{\theta} = \frac{\sum_i y_i}{n}$.

Improper Prior

This brings us to the subject of improper priors.

An improper prior is one in which the prior density may not integrate to unity over its domain.

It is sometimes possible define an improper prior in a way that makes it possible compute, in the usual Bayesian way, a proper posterior

$$P(\theta|y) \propto p(y|\theta)p(\theta)$$

Normal data with a noninformative prior distribution

By way of example let's consider estimating the mean of a population from a sample (given a vector y of n iid observations from a univariate normal distrbution, $N(\mu, \sigma^2)$.

Note that in this case we might consider σ^2 a "nuisance parameter" since we don't care about it's value. So, although we will construct a proper posterior distribution over

$$P(\mu, \sigma^2 | y) \propto P(y | \mu, \sigma^2) P(\mu, \sigma^2)$$

(using an improper prior) what we are really interested in is

$$P(\mu | y) = \int P(\mu, \sigma^2 | y) d\sigma^2$$

Foreshadowing the regression application of this, consider predicting the mean response for a given input where we don't know the variance.

Normal data with a noninformative prior distribution

Consider the prior

$$P(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

This yields a posterior

$$\begin{aligned} P(\mu, \sigma^2 | y) &\propto \sigma^{-n-2} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\ &= \sigma^{-n-2} exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]\right) \end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance of the y_i's.

Posterior distribution of $\mu|y$

After significant algebra (Gelman, pg. 76) it is possible express

$$P(\mu|y) = \int P(\mu,\sigma^2)d\sigma^2 \propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}$$

which is, by inspection, a student-t distribution $\theta \sim t_\nu(\psi,\rho^2)$ with $\nu = n-1$ degrees of freedom, mean $\psi = \bar{y}$ and scale $\rho^2 = \frac{s^2}{n}$. The pdf of the student-t distribution is given by

$$P(\mu|\psi,\nu,\rho) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\mu\pi\rho}}(1 + \frac{1}{\nu}(\frac{\mu-\psi}{\rho})^2)^{-(\nu+1)/2}$$

Relation to sampling theory

We have just shown that under the prior $P(\mu, \sigma^2) \propto (\sigma^2)^{-1}$ the posterior distribution of μ is given by

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

Under the sampling distribution $P(y|\mu, \sigma^2)$ the following relation holds when n is small.

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

Thus one way of interpreting a t-test about a population mean is as posterior inference in a Bayesian model with an improper (uniform over some domain) prior.

Posterior Predictive Inference

Using the same Bayesian machinery, the posterior predictive distribution

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

can be shown to be distributed according to a student-t distribution with location $\bar{y}$, scale $(1+\frac{1}{n})^{1/2}s$ and $n-1$ degrees of freedom.

Note: we are working towards a Bayesian explanation of regression so you should be thinking predicting the output for a given input while integrating out the unknown regression model parameters. The calculation (and results) for the normal regression model will be similar.