Outline

The Unknown μ Model

The model is

$$Y_i = \mu + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2), \; i = 1, 2, \ldots, n.$$

We will assume we know σ and want inference for μ.

The Conjugate Prior

The normal is conjugate:

$$\mu \sim N(\bar{\mu}, \tau^2).$$

The Likelihood

$$p(y_1, y_2, \ldots, y_n \mid \mu) \propto e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2}$$

$$\sum(y_i - \mu)^2 = \sum(y_i - \bar{y})^2 + n(\mu - \bar{y})^2$$

$$L(\mu) \propto e^{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2}$$

The Posterior

$$L(\mu) \;\times\; p(\mu)$$

$$\propto e^{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2} \;\times\; e^{-\frac{1}{2\tau^2}(\mu-\bar{\mu})^2}$$

$$= e^{-\frac{1}{2}(a(\mu-\bar{y})^2+b(\mu-\bar{\mu})^2)}$$

with

$$a = \frac{n}{\sigma^2}, \quad b = \frac{1}{\tau^2}$$

$$(a(\mu - \bar{y})^2 + b(\mu - \bar{\mu})^2)$$

$$= (a+b)\,\mu^2 - 2\mu\,(a\,\bar{y} + b\,\bar{\mu}) + C$$

where C does not have μ in it.

$$= (a+b)(\mu^2 - 2\mu\tfrac{(a\,\bar{y}+b\,\bar{\mu})}{a+b}) + C$$

Let

$$\tilde{\mu} = \frac{(a\,\bar{y} + b\,\bar{\mu})}{a+b}.$$

$$(a(\mu - \bar{y})^2 + b(\mu - \bar{\mu})^2)$$

$$= (a + b)(\mu - \tilde{\mu})^2 + \tilde{C}$$

$$p(\mu \mid y) \propto e^{-\frac{a+b}{2}(\mu - \tilde{\mu})^2}$$

$$\mu \mid y \sim N(\tilde{\mu}, \frac{1}{a+b}).$$

with

$$a = \frac{n}{\sigma^2}, \quad b = \frac{1}{\tau^2}, \quad \tilde{\mu} = \frac{(a\,\bar{y} + b\,\bar{\mu})}{a+b}.$$

Comments

Note:

You can think of a as the weight given to the data and b as the weight given to the prior.

As n gets big the data dominates, as τ gets small the prior dominates.

b is called the prior *precision*.

The posterior mean is a weighted combination of the sample mean and the prior mean.

Note:

If you let $\tau \to \infty$ to get

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

which looks like the frequentist approach.

If you use the "improper prior" $p(\mu) \propto 1$ you get the above.

This prior is often describe as " non-informative".

Don't do this!

Note:

We often use the posterior mean as our estimate of μ.

This is optimal under squared error loss.

For finite τ we "shrink" the posterior from the sample mean towards the prior mean, where the amount of shrinkage is determined by τ.

$L(\mu) \times p(\mu)$

$$\propto e^{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2} \times e^{-\frac{1}{2\tau^2}(\mu-\bar{\mu})^2}$$

If we take the log we get

$$log(p(\mu \mid y)) = -\frac{n}{2\sigma^2}(\mu - \bar{y})^2 - \frac{1}{2\tau^2}(\mu - \bar{\mu})^2$$

If we set $\bar{\mu} = 0$, the second term is called a "penalty term" by data-miners.

Note:

For finite τ, the posterior mean is always a biased estimator.

Good.

Prediction

Let X denote the "next X".

What is the predictive distribution of X?

Have

$$X \mid \mu \sim N(\mu, \sigma^2).$$

$$\mu \sim N(\bar{\mu}, \tau^2),$$

where, again, $\bar{\mu}$ and τ could correspond to the prior or posterior versions.

then

$$\begin{aligned} Y &= \mu + \epsilon,\ \epsilon \sim N(0, \sigma^2) \\ &\sim N(\bar{\mu}, \sigma^2 + \tau^2). \end{aligned}$$