STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 4A: Poisson Models

Combining Information: Posterior Precision

- The (conditional) posterior parameters τ_1^2 and μ_1 combine the prior parameters τ_0^2 and μ_0 with terms from the data.
- We already saw that

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- Now, let
  - $\tilde{\sigma}^2 = 1/\sigma^2$. This is sampling precision – how "close" the x_i's are to μ, as opposed to how dispersed they are.
  - $\tilde{\tau}_0^2 = 1/\tau_0^2$. This is prior precision. How sharp my prior beliefs are.
  - $\tilde{\tau}_1^2 = 1/\tau_1^2$. This is posterior precision. How sharp my posterior beliefs are.
- It is convenient to think about precision as the quantity of information on an additive scale. For the normal model, we can now express posterior precision as:

$$\tilde{\tau}_1^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2$$

- And so in that sense: posterior information = prior informaton + data information

Combining Information: Posterior Mean

- We can now write the posterior mean in terms of prior and posterior precision as follows:

$$\mu_1 = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\overline{x}$$

- The posterior mean is a weighted average of the prior mean and the sample mean, as it was in the beta binomial model.
- The weight on the sample mean is n/σ^2, the sampling precision of the sample mean.
- The weight on the prior mean is $1/\tau_0^2$, the prior precision.
- Example: If the prior mean were based on k_0 prior observations from the same (or a similar) population as $X_1, \ldots, X_n$ (i.e., $k_0 = \alpha + \beta$ in the binomial model), then we might want to set $\tau_0^2 = \sigma_0^2/k_0$, the variance of the mean of the prior observations. In this case, the formula for the posterior mean reduces to:

$$\mu_1 = \frac{k_0}{k_0 + n}\mu_0 + \frac{n}{k_0 + n}\overline{x}$$

Normal Predictions

- Suppose we want to predict $\widetilde{Y}$ from the population after observing $Y_1 = y = 1, ..., Y_n = y_n$.
- Note that $\widetilde{Y} \sim \mathrm{N}(\mu, \sigma^2) \leftrightarrow \widetilde{Y} = \mu + \widetilde{\epsilon}, \quad \widetilde{\epsilon} \sim \mathrm{N}(0, \sigma^2)$
- Saying that $\widetilde{Y}$ is normal with mean μ is the same as saying $\widetilde{Y}$ is equal to μ plus some mean-zero normally distributed noise.
- Using this, we can compute posterior mean and variance of $\widetilde{Y}$:

$$
\begin{aligned}
\mathrm{E}[\widetilde{Y}|y, \sigma^2] &= \mathrm{E}[\mu + \widetilde{\epsilon}|y, \sigma^2] \\
&= \mathrm{E}[\mu|y, \sigma^2] + \mathrm{E}[\widetilde{\epsilon}|y, \sigma^2] \\
&= \mu_1 + 0 = \mu_1 \\
\mathrm{Var}(\widetilde{Y}|y, \sigma^2) &= \mathrm{Var}(\mu + \widetilde{\epsilon}|y, \sigma^2) \\
&= \mathrm{Var}(\mu|y, \sigma^2) + \mathrm{Var}(\widetilde{\epsilon}|y, \sigma^2) \\
&= \tau_1^2 + \sigma^2
\end{aligned}
$$

- Recall that the sum of iid normal RVs is also normal.
- Hence, since both μ and $\widetilde{\epsilon}$, conditional on y and σ^2, are normal, so is $\widetilde{Y} = \mu + \widetilde{\epsilon}$.
- The predictive distribution is therefore

$$\widetilde{Y}|\sigma^2, y \sim \mathrm{N}(\mu_1, \tau_1^2 + \sigma^2)$$

Normal Predictions

- Our uncertainty about a new sample $\widetilde{Y}$ is therefore a function of our uncertainty about the center of the population as well as how variable the population is. As $n \to \infty$ we become more confident about where μ is. But certainty about μ does not reduce the sampling variability, and so our uncertainty about $\widetilde{Y}$ never goes below σ^2.

Normal Predictions

- Grogan and Wirth (1981) provide data on the wing length in millimeters of nine members of a species of midge. From these nine measurements we wish to make inferences on the population mean μ. Studies from other populations suggest that wing lengths are typically around 1.9 mm, and so we set $\mu_0 = 1.9$.
- Suppose we set $\tau_0 = 0.95$.
- The mean of the observations is $\overline{y} = 1.804$ and $s^2 = 0.017$.
- Find μ_1 and τ_1^2 and $\pi(\mu|y, s^2)$
- Find a 95% credible interval for μ based on $\pi(\mu|y, s^2)$.
- This example assumes that $s^2 = \sigma^2$. To get a more accurate representation of uncertainty we will need to develop a model where σ^2 is also unknown.

Unknown Variance, Known Mean

- Let $y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ where μ is unknown, σ^2 known.
- The likelihood function is

$$\pi(y|\sigma^2) = (2\pi\sigma^2)^{n/2} \exp\left[-(2\sigma^2)^{-1} \sum_{i=1}^{n} (y_i - \mu)^2 \right]$$

- For a natural conjugate prior, we seek a prior that has the same functional form:

$$\pi(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2} \right]$$

- This is the kernel of an inverse gamma distribution. If $y \sim \mathrm{G}^{-1}(\alpha, \beta)$ then

$$f(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (1/y)^{\alpha+1} \exp(-\beta/y)$$

- Thus the natural conjuate prior in this case is $\sigma^2 \sim \mathrm{G}^{-1}(\alpha, \beta)$
- The normalizing constant is $\frac{\beta^\alpha}{\Gamma(\alpha)}$

Unknown Variance, Known Mean

- The posterior distribution is given by

$$\pi(\sigma^2|y,\mu) \propto (\sigma^2)^{-(\alpha+1)} \exp\big[-\frac{\beta}{\sigma^2}\big](2\pi\sigma^2)^{n/2} \exp\big[-(2\sigma^2)^{-1}\sum_{i=1}^{n}(y_i-\mu)^2\big]$$
$$\propto (\sigma^2)^{-(n/2+\alpha+1)} \exp\left[-\frac{v/2+\beta}{\sigma^2}\right]$$

where $v = \frac{1}{n}\sum(y_i-\mu)^2$

- Hence $\sigma^2|y \sim \mathrm{G}^{-1}(n/2+\alpha, v/2+\beta)$

An Improper Prior

- For a $\mathrm{G}^{-1}(\alpha, \beta)$ prior, as $\alpha \downarrow 0$ and $\beta \downarrow 0$ an improper prior results, as it did in the case of a $\mathrm{Beta}(\alpha, \beta)$ prior. In this case,

$$\lim_{\alpha \to 0, \beta \to 0} \pi(\sigma^2 | \alpha, \beta) = \sigma^{-2}$$

- The posterior distribution is then

$$\sigma^2 | y \sim \mathrm{G}^{-1}(n/2, v/2)$$

- This distribution is proper

Precision

- Remember we said in Bayesian approaches we often like to work directly with the precision σ^{-2}. as opposed to the variance.
- The likelihood can be written as

$$\pi(y|\sigma^{-2}) \propto (\sigma^{-2})^{n/2} \exp\left[-\frac{1}{2}\sigma^{-2}\sum(y_i-\mu)^2 \right]$$

- Using the fact that if $y \sim \mathrm{G}^{-1}(\alpha,\beta)$ then $y^{-1} \sim \mathrm{G}(\alpha,\beta)$ it follows that the natural conjugate prior for σ^{-2} is

$$\sigma^{-2} \sim \mathrm{G}(\alpha,\beta)$$

- Therefore,

$$\pi(\sigma^{-2}|y) \propto (\sigma^{-2})^{\alpha+n/2-1} \exp[-(\beta+v/2)\sigma^{-2}]$$

- Hence

$$\sigma^{-2}|y \sim \mathrm{G}(\alpha+n/2, \beta+v/2)$$

- As $\alpha \downarrow 0, \beta \downarrow 0$, an improper prior results

$$\pi(\sigma^{-2}) = \sigma^2$$

Poisson Distribution

- Poisson distribution models the probability of events occurring in a fixed interval, provided they occur at a constant rate and independently of the time since the last occurrence.
- This distribution is frequently used in operations management and supply chain logistics.
- For example: A distribution center receives 180 orders per hour; 24 hours a day. The orders are independent in the sense that receiving one does change the probability of when the next one will arrive.
- Our sample space is $\{0, 1, 2, ...\}$.
- Notice that this can be used to model any situation where the measurements are non-negative whole numbers, provided the rate of occurrence is constant and the probability of one occurrence is independent of the previous.
- A random variable X has a Poisson distribution with mean θ if

$$\Pr(X = k|\theta) = \frac{\theta^k}{k!}\exp(-\theta) \;\; \text{for } k \in \{0, 1, 2, ...\}$$

- The mean and variance of X are both θ.
- Hence the Poisson family of distributions has a mean-variance relationship in the sense that if one Poisson distribution has a larger mean than another, it will have a larger variance as well.

Conjugate Prior

- Suppose we have n iid Poission observations $y = (y_1, \ldots, y_n)$ with mean θ, then the likelihood is given by

$$\begin{aligned}\pi(y|\theta) &= \prod_i \pi(y_i|\theta) \\ &= \prod_i \frac{1}{y_i!}\theta^{y_i}\exp(-\theta) \\ &\propto \theta^{\sum y_i}\exp[-n\theta] \\ &\propto \exp[a(y)\theta + b(y)\log(\theta)]\end{aligned}$$

  where $a(y) = -n$, and $b(y) = \sum y_i$, (using that $e^{k\log(p)} = p^k$).
- Note also that $\sum Y_i$ is a sufficient statistic for θ and $\sum Y_i \sim \text{Poisson}(n\theta)$.

Conjugate Prior

- We know that our posterior has the following form

$$\pi(\theta|y) \propto \pi(\theta) \times \theta^{\sum y_i} \exp(-n\theta)$$

- This means that whatever our conjugate class of densities is, it will have to include terms like $\theta^{c_1} \exp(-c_2\theta)$ for numbers c_1, c_2.
- The simplest class of such densities includes only these terms, and their corresponding probability distributions are known as the family of gamma distributions.
- Hence, $\theta \sim \mathrm{G}(\alpha, \beta)$ if

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp[-\beta\theta] \quad \text{for } \theta, \alpha, \beta > 0$$

- The mean is given by $\mathrm{E}[\theta] = \alpha/\beta$, and the variance is $\mathrm{Var}(\theta) = \alpha/\beta^2$
- This is a gamma distribution with $\alpha > 0$ shape and $\beta > 0$ rate. Be careful because sometimes the gamma distribution is parameterized in terms of shape and a scale parameter.

Posterior Distribution

- Let $y_i|\theta \sim \text{Poisson}(\theta)$ and $\theta|\alpha, \beta \sim \text{G}(\alpha, \beta)$.
- Then the posterior distribution is given by:

$$\begin{aligned}\pi(\theta|y) &\propto \pi(\theta)\pi(y|\theta) \\ &\propto \left(\theta^{\alpha-1}\exp[-\beta\theta]\right) \times \left(\theta^{\sum y_i}\exp[-n\theta]\right) \\ &\propto \theta^{\alpha+\sum y_i - 1}\exp[-(\beta+n)\theta]\end{aligned}$$

- This is recognizable as a gamma distribution with parameters $\alpha + \sum y_i, \beta + n$.
- Hence, if $\theta \sim \text{G}(\alpha, \beta)$ and $Y_1, ...Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ then

$$\theta|Y_1, ...Y_n \sim \text{G}(\alpha + \sum Y_i, \beta + n)$$

Posterior Inference

- Estimation and prediction proceed in a manner similar to that in the binomial model.
- The posterior mean of θ is a convex combination of the prior mean and the sample mean:

$$\mathrm{E}[\theta|y] = \frac{\alpha + \sum y_i}{\beta + n} = \frac{\beta}{\beta + n}\frac{\alpha}{\beta} + \frac{n}{\beta + n}\frac{\sum y_i}{n}$$

- β is interpreted as the number of prior observations,
- α is interpreted as the sum of counts from b prior observations.
- For large n the information from the data dominates the prior information:

$$n >> \beta \rightarrow \mathrm{E}[\theta|y] \approx \overline{y}, \mathrm{Var}(\theta|y) \approx \overline{y}/n$$

Posterior Inference Summary

$$
\begin{aligned}
E(\theta) &= \alpha\beta^{-1} & E(\theta \mid y) &= (\alpha + n\bar{y})(\beta + n)^{-1} \\
\mathrm{Var}(\theta) &= \alpha\beta^{-2} & \mathrm{Var}(\theta \mid y) &= (\alpha + n\bar{y})(\beta + n)^{-2} \\
\mathrm{Mode}(\theta) &= (\alpha - 1)\beta^{-1} & \mathrm{Mode}(\theta \mid y) &= (\alpha - n\bar{y} - 1)(\beta + n)^{-1}
\end{aligned}
$$

Predictive Distribution

- Predictions about additional data can be obtained as before with the posterior predictive distribution, which is given by

$$\begin{aligned}
\pi(\widetilde{y}|y) &= \int_0^\infty \pi(\widetilde{y}|\theta, y)\pi(\theta|y)d\theta \\
&= \int \pi(\widetilde{y}|\theta)\pi(\theta|y)d\theta \\
&= \int \widetilde{y} \sim \text{Poisson}(\theta) \times \theta \sim \text{G}(\alpha + n\overline{y}, \beta + n)d\theta \\
&= \frac{(\beta+n)^{\alpha+\sum y_i}}{\Gamma(\widetilde{y}+1)\Gamma(\alpha+\sum y_i)} \int_0^\infty \theta^{\alpha+\sum y_i+\widetilde{y}-1} \exp[-(\beta+n+1)\theta]d\theta
\end{aligned}$$

- This looks daunting but we can simplify it.
- Using what you know about the Gamma density, see if you can simplify it.

Predictive Distribution

$$1 = \int_0^\infty \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \, d\theta \quad \text{for any values } a, b > 0 \; .$$

This means that

$$\int_0^\infty \theta^{a-1} e^{-b\theta} \, d\theta = \frac{\Gamma(a)}{b^a} \quad \text{for any values } a, b > 0 \; .$$

Now substitute in $a + \sum y_i + \tilde{y}$ instead of a and $b + n + 1$ instead of b to get

$$\int_0^\infty \theta^{a + \sum y_i + \tilde{y} - 1} e^{-(b+n+1)\theta} \, d\theta = \frac{\Gamma(a + \sum y_i + \tilde{y})}{(b+n+1)^{a + \sum y_i + \tilde{y}}} \; .$$

After simplifying some of the algebra, this gives

$$p(\tilde{y} | y_1, \ldots, y_n) = \frac{\Gamma(a + \sum y_i + \tilde{y})}{\Gamma(\tilde{y}+1)\Gamma(a + \sum y_i)} \left(\frac{b+n}{b+n+1} \right)^{a + \sum y_i} \left(\frac{1}{b+n+1} \right)^{\tilde{y}}$$

Predictive Distribution

- This is a negative binomial distribution with parameters $(\alpha + \sum y_i, \beta + n)$, and the posterior mean and variance are given by:

$$
\begin{aligned}
\mathrm{E}[\tilde{Y}|y_1,\ldots,y_n] &= \frac{a+\sum y_i}{b+n} = \mathrm{E}[\theta|y_1,\ldots,y_n]; \\
\mathrm{Var}[\tilde{Y}|y_1,\ldots,y_n] &= \frac{a+\sum y_i}{b+n}\frac{b+n+1}{b+n} = \mathrm{Var}[\theta|y_1,\ldots y_n] \times (b+n+1) \\
&= \mathrm{E}[\theta|y_1,\ldots,y_n] \times \frac{b+n+1}{b+n}.
\end{aligned}
$$