Bioassay Example: Completing the Picture

In class we considered the bioassay where a compound was given at four log-dosage levels (x_i) to different groups of animals (n_i animals for each dosage) and where in each group y_i animals died. We used θ_i to denote the probability of death for log-dosage x_i and after some discussion decided on a logistic regression model

$$logit(\theta_i) \equiv \log \frac{\theta_i}{1-\theta_i} = \alpha + \beta x_i$$

Assuming a binomial model given θ_i and independence of observations given the parameters we have

$$p(y \mid \alpha, \beta, n, x) = \prod_{i=1}^{k} \left[logit^{-1}(\alpha + \beta x_i)\right]^{y_i} \left[1 - logit^{-1}(\alpha + \beta x_i)\right]^{n_i - y_i}$$

Choosing a uniform prior $p(\alpha, \beta) \propto 1$ (or a more complex one if we later deem this too naive or have an informative prior), we now face the problem of computing the posterior

$$p(\alpha, \beta \mid y, n, x) \propto p(\alpha, \beta) p(y \mid \alpha, \beta, n, x) \tag{1}$$

To do so we will use the following computational (in contrast to analytical) approach

1. Roughly estimate the effective range of α and β
2. Compute the unnormalized posterior distribution on a discrete grid
3. Approximate the posterior by normalizing over the grid
4. Draw 1000 samples from the posterior distribution
5. Use the 1000 samples to compute quantities of interest

I will now briefly describe each of these steps.

1. **Rough estimate of the parameters**. To obtain a rough estimate of the parameters we note that by our choice of the model $logit(E[y_i/n_i \mid \alpha, \beta]) = \alpha + \beta x_i$. We can thus crudely estimate the parameters by a linear regression (or a logistic regression) of $logit(y_i/n_i)$ on x_i on the four data points. The linear regression estimate is $(\hat{\alpha}, \hat{\beta}) = (0.1, 2.9)$ with standard errors of 0.3 and 0.5.

2. **Compute the unnormalized posterior distribution on a grid**. Using the above rough estimate, we can start by constructing a discrete 200×200 grid over the approximate range of two standard errors $(\alpha, \beta) \in [-1, 1] \times [1, 5]$. For each point on this grid we can easily evaluate the unnormalized posterior using Eq. (1) and draw a contour plot (lines of equi-probability) of the distribution as in the left-hand figure in the handout given in class (there is also a link to it in the website). In fact, the first attempt at doing this resulted in $5 - 95\%$ contour lines that were outside of the crude range and so the graph is actually the result of expanding the range to $(\alpha, \beta) \in [-5, 10] \times [-10, 40]$.

3. **Normalizing the distribution**. We now use the grid values to approximate the normalized distribution. That is, we treat each point in the grid as covering a $d \times d$ square centered around the grid point (where d is the distance between points in the grid). To have the integral over the grid sum to one we simply divide each value in the grid by the sum of all values times d^2. Similarly, if we want to compute the posterior distribution over α alone, we first sum over the values of β and then divide by the sum of all value times d.

4. **Sampling from the posterior distribution**. We start by computing the marginal posterior distribution over α as described above and then create 1000 samples by repeating the following steps for each $l = 1, \ldots, 1000$:

   (a) draw α^l from the distribution $p(\alpha \mid y)$ we computed.

   (b) draw b_l from the discrete conditional distribution $p(\beta \mid \alpha^l, y)$ (computed using the same technique as above for each grid value of α).

   (c) For both α^l and β^l add a uniform jitter centered at zero and with width d. This makes the sampling distribution continuous.

   The right-hand figure in the handout shows the 1000 samples created in this manner. It is easy to see that its mass corresponds to the unnormalized density of the left-hand figure.

5. **Using the posterior samples**. We can now easily use the 1000 to compute quantities of interest. For example, the dosage level LD50 at which the probability of death is 50% is often of interest in bioassay studies. In our model, this means:

$$LD50 : E\left(\frac{y_i}{n_i}\right) = logit^{-1}(\alpha + \beta x_i) = 0.5$$

   from which we get that the LD50 is $x_i = -\alpha/\beta$. We compute this quantity for all 1000 samples and show a histogram of these posterior LD50 values on the bottom figure in the handout.