Bayesian inference 2017
Exercise session 6 (11.–14.12.2017)

1. Continued from the exercise 1 of the previous week. Show that the marginal posterior for the parameter μ is:

$$p(\mu|y) \propto \left(1 + \frac{\kappa_n(\mu - \mu_n)^2}{\nu_n \sigma_n^2}\right)^{-(\nu_n+1)/2},$$

that is, the marginal posterior for μ is the non-standard t-distribution $t_{\nu_n}(\mu_n, \sigma_0^2/\kappa_n)$.

2. Derive the posterior for the complete pooling model of Section 6.3.2 of the lecture notes. Assume normally distributed independent observations $Y_1, \ldots, Y_J$ with the same unknown mean θ and different, but known variances $\sigma_1, \ldots, \sigma_J$, and an improper uniform prior for the mean θ, so that the model is:

$$\begin{aligned} Y_j &\sim N(\theta, \sigma_j^2) \quad \text{for all } j = 1, \ldots, J \\ p(\theta) &\propto 1 \end{aligned}$$

Show that the posterior for the parameter θ is

$$p(\theta|\mathbf{y}) = N\left(\frac{\sum_{j=1}^J \frac{1}{\sigma_j^2} y_j}{\sum_{j=1}^J \frac{1}{\sigma_j^2}}, \frac{1}{\sum_{j=1}^J \frac{1}{\sigma_j^2}}\right).$$

3. Instead of the boring toy examples, let's solve a proper gambling example for the last exercise of the course (disclaimer: lecturer does not endorse gambling in any form, and especially not aided by dubious statistical models)!

File `football2017.csv` includes some results of the football matches from last year, along with the estimated probabilities for the home win, draw, and away win (variable names should be self-explanatory).

We will assume that the probability estimates supplied are more or less correct. But because it would be pretty boring to use these probability estimates to predict the winner of the match, let's use them to predict the final scores of the match instead.

We will make a bit unrealistic assumptions that the home and away scores are independent. This means that you can make a new data set by concatenating `y <- c(goalsHomeTeam, goalsAwayTeam)` and `x <- c(home_prob, away_prob)`, so that you have only one predicted variable Y, which is the goals of the team, and one predictor x, which is the win probability of the team. The size of the data set n is 2 times number of matches.

Because the final scores of the teams, denoted by the $Y_1, \ldots, Y_n$ are non-negative integers, it is natural to model the with Poisson distribution. We will predict the mean of the Poisson distribution with a linear regression equation

$$\alpha + \beta x_i,$$

where $x_1, \ldots, x_n$ are the probabilities of the team winning (these are assumed as constants, so you do not need any prior for them). But because the parameter of the Poisson distribution must be positive, this regression equation must be exponentiated to constrain the parameter to the positive real axis. We will also use uniform improper priors for the regression coefficients:

$$\begin{aligned} Y_i \,|\, \alpha, \beta &\sim \text{Poisson}(e^{\alpha+\beta x_i}) \quad \text{for all } i = 1, \ldots, n \\ \alpha &\propto 1, \ \beta \propto 1. \end{aligned}$$

Fit this model with Stan, raport the posterior means of the exponents of the regression coefficients, and draw histogram of the marginal posterior of the exponent of the regression coefficient e^{β}.

4. Let's move on to the actual gambling part! We will compute the optimal bets for the Stuttgart - Leverkusen Bundes-league match. The file `stuttgart_leverkusen.csv` contains the net odds for some of the scores of this match. It contains the following columns:

1. `homeTeamGoals` : score of home team; denote these as i
2. `awayTeamGoals` : score of away team; denote these as j
3. `winshare` : win multiplier for the this score; for instance, if win share for score 1-1 were 3.90, and you had bet for it 5 euros, you would win $3.90 \cdot 5 = 19.50$ euros if this were the final score of the match. Denote these as $w_{i,j}$ (win multiplier for score $i - j$).

So we have to predict the probabilities for the two new independent observations from the same distribution: goals of Stuttgart, denoted as $\tilde{Y}_1$, and goals of Leverkusen, denoted as $\tilde{Y}_2$. Assume that the probabilities for the home win, draw, and away win are 0.30, 0.27 and 0.43, respectively, so that the values of the predictors for the new observations are $\tilde{x}_1 = 0.30$ and $\tilde{x}_2 = 0.43$.

(a) Using your model fitted in the previous exercise, compute the probabilities for all the home and away scores for the scores from 0 to 6. That is, compute the values of the posterior predictive distributions $p(\tilde{y}_1 = i \,|\, \mathbf{y})$ and $p(\tilde{y}_2 = i \,|\mathbf{y})$ for all $i = 0, \ldots, 6$.

You can approximate these probabilities using the Monte Carlo sample $(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(S)}, \beta^{(S)})$ generated by Stan:

$$p(\tilde{y}_1 = i \,|\, \mathbf{y}) \approx \frac{1}{S} \sum_{s=1}^{S} p(\tilde{y}_1 = i \,|\, \alpha^{(s)}, \beta^{(s)}).$$

(b) Now we can compute the expected values for the bets. Because we assumed the scores of the home and away teams independent, you can compute the probability of the score (i, j) by simply multiplying the probabilities for the home and away scores:

$$p(\tilde{y}_1 = i, \tilde{y}_2 = j \,|\, \mathbf{y}) = p(\tilde{y}_1 = i \,|\, \mathbf{y}) p(\tilde{y}_2 = j \,|\, \mathbf{y}).$$

Denote the winnings for the bet of one unit (for example, 1 euro) for score $i - j$ as a random variable

$$Z_{i,j} = \begin{cases} w_{i,j} & \text{if } \tilde{Y}_1 = i \text{ and } \tilde{Y}_2 = j, \\ 0 & \text{otherwise.} \end{cases}$$

Using the probabilities of the scores computed above, compute the expected winnings for the unit bet for all the scores listed on the table `stuttgart_leverkusen.csv`:

$$E(Z_{i,j} \,|\, \mathbf{y}) = w_{i,j}\, p(\tilde{y}_1 = i, \tilde{y}_2 = j \,|\, \mathbf{y}).$$

(c) Finally we will compute the optimal bets for this match! The Kelly criterion[1] states that the optimal fraction of your bankroll to bet for the outcome is

$$f^* = \frac{p(b+1) - 1}{b},$$

[1] `https://en.wikipedia.org/wiki/Kelly_criterion`

where p is the probability of the outcome, and b are the net odds[2] (computed as $w-1$, where w is the win multiplier). for the outcome. So in this match the optimal bet[3] for the score $i-j$ is

$$f^*_{i,j} = \frac{p(\tilde{y}_1 = i, \tilde{y}_2 = j \,|\, \mathbf{y}) w_{i,j} - 1}{w_{i,j} - 1}.$$

Assume that your bankroll is quite meager 1000 euros. Using the formula above, compute your optimal bets for this match (Kelly's criterion gives negative bets for the outcomes with negative expected value. Unfortunately we cannot place a negative bet, so we will not bet on these outcomes).

You can make the simplifying assumptions that our bets do not affect the net odds, and that there is no minimum bet size or other limits for the bets, so that you can place any bet (even bet of some arbitrary fractional number, such as 0.34 euros).

Sort the scores both according to the expected winnings of unit bet (so that the score with the largest expected value comes first) computed in (b), and to the sizes of the optimal bets (so that the score with the largest bet comes first) computed using the Kelly criterion. How would you explain the difference between the orders[4]?

[2]Net odds are the net profit for the unit bet: for instance, if the win multiplier were $w = 3.90$, the net profit for unit bet would be $b = w - 1 = 2.90$. This means that your net profit from a bet of 5 euros would be $5 \cdot 2.90 = 14.50$ euros.

[3]This is again little bit of the simplification: because we are betting on the multiple exclusive outcomes (multiple scores for one match), we should actually use the *generalized Kelly criterion*: `https://en.wikipedia.org/wiki/Kelly_criterion#Multiple_horses`. If you are feeling that this exercise was too easy, you can try using the generalized Kelly criterion instead to see how it affects the optimal bets!

[4]The match was played on Friday, so unfortunately you cannot actually place any bets anymore! Also the net odds were not final, so most probably it would not have been possible to gain any positive expected value from this match :)