Module 11: Linear Regression

Rebecca C. Steorts

Linear Regression Applied to Swimming

- We will consider Exercise 9.1 in Hoff very closely to illustrate linear regression.
- The data set we consider contains times (in seconds) of four high school swimmers swimming 50 yards.
- There are 6 times for each student, taken every two weeks.
- Each row corresponds to a swimmer and a higher column index indicates a later date.

Data set

```r
read.table("data/swim.dat",header=FALSE)
```

```
## Warning in read.table("data/swim.dat", header = FALSE): incomplete final line
## found by readTableHeader on 'data/swim.dat'
```

```
##     V1   V2   V3   V4   V5   V6
## 1 23.1 23.2 22.9 22.9 22.8 22.7
## 2 23.2 23.1 23.4 23.5 23.5 23.4
## 3 22.7 22.6 22.8 22.8 22.9 22.8
## 4 23.7 23.6 23.7 23.5 23.5 23.4
```

Full conditionals (Task 1)

We will fit a separate linear regression model for each swimmer, with swimming time as the response and week as the explanatory variable. Let $Y_i \in \mathbb{R}^6$ be the 6 recorded times for swimmer $i = 1, 2, 3, 4$. Let

$$X_i = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ ... & \\ 1 & 9 \\ 1 & 11 \end{bmatrix}$$

be the design matrix for swimmer $i = 1, 2, 3, 4$. Then we use the following linear regression model:

$$\begin{aligned} Y_i \mid \beta_i, \tau_i &\sim \mathcal{N}_6\left(X\beta_i, \tau_i^{-1}\mathcal{I}_6\right) \\ \beta_i &\sim \mathcal{N}_2\left(\beta_0, \Sigma_0\right) \\ \tau_i &\sim \text{Gamma}(a, b). \end{aligned}$$

Derive full conditionals for β_i and τ_i. Assume that β_0, Σ_0, a, b are known.

Solution (Task 1)

The conditional posterior for β_i is multivariate normal with

$$\begin{aligned}
\mathbb{V}[\beta_i \,|\, Y_i, X_i, \tau_i] &= (\Sigma_0^{-1} + \tau X_i^T X_i)^{-1} \\
\mathbb{E}[\beta_i \,|\, Y_i, X_i, \tau_i] &= (\Sigma_0^{-1} + \tau_i X_i^T X_i)^{-1}(\Sigma_0^{-1}\beta_0 + \tau_i X_i^T Y_i).
\end{aligned}$$

while

$$\tau_i \,|\, Y_i, X_i, \beta \sim \text{Gamma}\left(a + 3\,,\, b + \frac{(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)}{2}\right).$$

These can be found in in Hoff in section 9.2.1.

Task 2

Complete the prior specification by choosing a, b, β_0, and Σ_0. Let your choices be informed by the fact that times for this age group tend to be between 22 and 24 seconds.

Solution (Task 2)

Choose $a = b = 0.1$ so as to be somewhat uninformative.

Choose $\beta_0 = [23 \;\; 0]^T$ with

$$\Sigma_0 = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}.$$

This centers the intercept at 23 (the middle of the given range) and the slope at 0 (so we are assuming no increase) but we choose the variance to be a bit large to err on the side of being less informative.

Gibbs sampler (Task 3)

Code a Gibbs sampler to fit each of the models. For each swimmer i, obtain draws from the posterior predictive distribution for y_i^*, the time of swimmer i if they were to swim two weeks from the last recorded time.

Posterior Prediction (Task 4)

The coach has to decide which swimmer should compete in a meet two weeks from the last recorded time. Using the posterior predictive distributions, compute $\Pr\{y_i^* = \max(y_1^*, y_2^*, y_3^*, y_4^*)\}$ for each swimmer i and use these probabilities to make a recommendation to the coach.

- This is left as an exercise.