ST740 – Assignment 1 – Due 9/13

The homework is due in class on Wednesday, 9/13. Show work for questions that say "derive" and include code used to produce the results. You many work in groups, but please write up separate solutions.

1. Consider repeated measures linear regression with n subjects $i \in \{1, ..., n\}$ and m_i observations for subject i. Assume the linear model

$$Y_{ij} = \mathbf{X}_{ij}^T \beta + \varepsilon_{ij}$$

where $\mathbf{X}_{ij} = (X_{ij1}, ..., X_{ijq})^T$ are known covariates and $\beta = (\beta_1, ..., \beta_q)^T$ are the unknown regression coefficients. The errors are independent across subject i and correlated within subject as $(\varepsilon_{i1}, ..., \varepsilon_{im_i})^T \sim \text{Normal}(0, \Sigma_i)$ for $m_i \times m_i$ known covariance matrix Σ_i whose inverse is denoted $W_i = \Sigma_i^{-1}$.

   (a) Select a proper conjugate prior distribution for β and derive its posterior distribution.

   (b) Verify using a few simple special cases that your derivation in (a) is correct.

   (c) Propose a diagnostic to determine which subjects provide the most information about β.

2. Download the `lung` data from the `survival` package in `R`. Let Y_i be the survival time (days) for subject $i \in \{1, ..., n\}$. For this assignment, we will make the unrealistic assumption that the survival times follow an exponential distribution,

$$Y_i|\lambda \sim \text{Exponential}(\lambda).$$

The study runs from time 0 to time T, with patient i joining the study at time $T - T_i$ so that they are part of the study for T_i days. The analysis is complicated by censoring, i.e., some patients have not experienced their event at the conclusion of the study. An event is observed if $Y_i < T_i$ and censored if $Y_i > T_i$, with $\delta_i = I(Y_i > T_i)$ denoting the binary indicator that the event is censored.

For uncensored patients with $\delta_i = 0$ the contribution to the likelhood function is the exponential PDF, $\lambda \exp(-\lambda y_i)$. For censored patients with $\delta_i = 1$ the contribution to the likelihood function is the probability that they survive beyond T_i days, $\exp(-\lambda T_i)$.

   (a) Write the likelihood function in terms of parameter λ and data $(Y_1, T_1, \delta_1), ..., (Y_n, T_n, \delta_n)$.

   (b) Derive the Jeffreys prior for λ assuming Y_i is random but T_i is fixed, and give conditions that ensure it is a proper distribution. Plot this prior for the lung data and comment on whether it is uninformative.

   (c) Derive a conjugate prior distribution for λ are propose hyperparameters that lead to an uninformative prior distribution.

   (d) Using the full dataset, summarize the posterior distribuiton of λ under a few conjugate priors (i.e., different hyperparameters). Are the results senstive to the prior?

   (e) Using an uninformative conjugate prior, analyze the data separately for men and women and test whether the survival distribution varies by sex.