L10 Model evaluation and comparison

1. Model-checking or evaluation, that is measuring model goodness-of-fit and predictive ability, is an important component of Bayesian data analysis that we have not discussed much yet.

   (a) One method to assess goodness-of-fit is to compare posterior predictions of replicated data with the observed data. If the model fits well, then replicated data generated under the model should look like the observed data.

   (b) Let y^{rep} be replicated data that could have been observed, or data we would see tomorrow if the experiment that produced y today were replicated with the same model and parameter values that produced y.

   (c) The posterior predictive distribution is $P(y^{\text{rep}}|y) = \int P(y^{\text{rep}}|\theta)P(\theta|y)d\theta$.

   (d) Another method to assess goodness-of-fit that is particularly useful to evaluate constant variance assumption is to generate residual plots.

   (e) $r_i = y_i - \mu_i$ is the residual, or difference between the observed and expected value.

   (f) Lets apply these approaches to the guinea pig tooth length data, recall the data:

   `ToothGrowth` data set in `R`. The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

   $y_i \sim \text{Normal}(\mu_i, \tau)$
   $\mu_i = \beta_1 \mathbb{1}_{OJ} + \beta_2 \mathbb{1}_{VC} + \beta_3 x_{\text{dose}}$
   $\beta_i \sim \text{Normal}(\mu_0 = 0, \tau_0 = 1e^{-6})$
   $P(\beta, \tau|\mathbf{y}, \mathbf{x}) \propto \prod_i P(y_i|\beta, \mathbf{x}, \tau) \prod_j P(\beta_j)P(\tau)$

   (g) See the `JAGS` model and `R` code for this example: `toothlmModelV2` and `tooth.R`. Four ways of assessing model-fit are included in `tooth.R`.

      i. Posterior predictive check: compares the predictive distribution y^{rep} to the observed data y.
      ii. Residual plot: examines residuals as a function of the mean.
      iii. Bayesian r^2 analogue: measures the proportional reduction of uncertainty concerning the response variable y achieved by incorporating the explanatory variables x in the model.
      iv. Cross-validation posterior predictive check: similar to the posterior predictive check, but does not involve double use of the data.

   (h) Another example, daily air quality measurements in New York, May to September 1973. The response we are interested in is Ozone readings. We are interested in whether Ozone varies as a function of wind speed or temperature, and whether there is an interaction between wind-speed and temperature.

i. The data can be obtained using `data(airquality)`. There is some missing data denoted `NA` don't worry about this too much. But when you run the MCMC include the data as one of the variables you monitor.

ii. Here is the sampling distribution and linear model:

$y_i \sim \text{Normal}(\mu_i, \tau)$
$\mu_i = \beta_1 + \beta_2 x_w + \beta_3 x_T + \beta_4 x_w x_T$

iii. Work in groups to write, compile, and run the model using `JAGS`. The files `airmodel` and `airquality.R` can be used to get you started, look at these if they are stuck. Once you run the model first evaluate mixing, then model-fit, and then (if everything looks ok) examine the posterior to answer the question posed above.

2. Model choice and model comparison using the deviance information criterion

(a) Often in biology we wish to compare alternative models. There are several ways to do this, e.g., DIC, Bayes Factor, model posterior probabilities. Bayesian model averaging is an alternative method that incorporates model uncertainty in parameter estimates. We will focus on DIC because it is widely used and easy to calculate.

(b) The deviance information criterion (DIC) is a hierarchical modeling generalization of the Akaike information criterion and is easily calculated from MCMC output. The deviance is proportional to the likelihood,
$D(\theta) = -2log(P(y|\theta)) + C$, where $p(y|\theta)$ is the model likelihood and C is a constant that we dont need to know or worry about.

(c) We can obtain a point estimate of a models deviance by using the parameter point estimates to estimate deviance, i.e., $D(\bar{\theta}) = D(\bar{\theta}, y)$. This does not account for uncertainty, but is a necessary step in calculating DIC.

(d) Next we must calculate the expected deviance, which is an overall measure of model fit. We compute this by averaging over the posterior distribution.
$\bar{D} = E[D(\theta, y)] = \int D(\theta, y) P(\theta|y) d\theta$

(e) Model fit should be penalized by the number of parameters, we use the effective number of parameters, which should generally be less than or equal to the true number of parameters. The effective number of parameters is lower when parameters are not independent. We calculate the effective number of parameters as,
$pD = \bar{D} - D(\bar{\theta})$.

(f) We then calculate DIC as,
$DIC = \bar{D} + pD$
Lower values suggest a better model and models are penalized by the effective number of parameters, this is a penalty for model complexity. The model with the lowest DIC is the best model (of those compared), but there is not an ideal

way to assess the significance of differences in DIC. This same issue applies to AIC. Other solutions to model comparison exist (bayes factor and BMA).

(g) tt dic.samples is used to estimate DIC for each model. It returns $\bar{D}$ and pD for each stochastic node, not the model as a whole. The `difdic` function is used to estimate the diffrence in DIC between models.