7 Week 7

7.1 Posterior predictive check: Mauna Loa CO_2 continued

R-template `ex_linearRegression_postcheck.R`.

In the original Mauna Loa CO_2 data analysis we visualized the posterior predictive distribution of the expected CO_2 with respect to the month and compared it to the observed data points. This can be seen as one method for visual posterior predictive check. However, let's continue model checking a bit more and then improve the model based on our findings.

Conduct posterior predictive check for the Mauna Loa CO_2 data in similar manner as in the Speed of Light example in BDA3. Sample 20 replicates of the *data set* and do the following:

1. Plot histograms of the replicate data sets. Compare the histograms of replicate data sets to the histogram of the real data. Discuss whether the replicate histograms look similar to the real data histogram – remember to justify your discussion.

To sample a replicate data set you must sample $\tilde{y} = [\tilde{y}_1, \ldots, \tilde{y}_n]$ values from $\tilde{y}_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = a + bx_i$ and a, b, σ^2 are drawn from the posterior. For example, pick 20 random triplets of a, b, σ^2 from the Markov chain and for each of them sample $\tilde{y}$. Then plot histogram of each $\tilde{y}$.

Next, revise the model so that $\mu_i = a + bx_i + cx_i^2$. Find the posterior of the parameters of the new model and do the following:

2. Plot the posterior mean and central 95% credible interval of μ and $\tilde{y}$ as a function of x with the new model. Overlay this plot with the data. Is there visual improvement in the fit between the 95% credible intervals and observations? If yes, how?

3. Do the same full data posterior predictive check as with the original model by visualizing the histograms of the full true data and 20 full replicate data sets. Discuss whether the replicate histograms look similar to the real data histogram – remember to justify your discussion. Did the model refinement improve models behaviour in this respect?

Note! Since you are not asked about the parameter inference, you don't need to worry about how to scale $\dot{c}$ back to c even if you standardize your y and x.

The danger with sequantial model refinements is that we conduct it so long that our model overfits the data. Hence,

4. compare these two alternative models (M1: $\mu_i = a + bx_i$, M2: $\mu_i = a + bx_i + cx_i^2$) with posterior predictive comparison by dividing the data into two parts and taking every other observation into training data and every other observation into test data.

Conduct the posterior predictive comparison using the point-wise log predictive density

$$\text{lpd} = \sum_{i=1}^{n_{\text{test}}} \log p(\tilde{y}_i | \tilde{x}_i, y_{\text{training}}, x_{\text{training}})$$

and the root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\text{E}[\tilde{y}_i | \tilde{x}_i, y_{\text{training}}, x_{\text{training}}] - \tilde{y}_i)^2}$$

where $y_{\text{training}}, x_{\text{training}}$ are the training and test data and $\tilde{y}_i, \tilde{x}_i$ are the test data points. Which of the models has better posterior predictive performance. Based on this results, does it seem that model M2 has overfitted the data?

Note,

$$\begin{aligned} p(\tilde{y}_i | \tilde{x}_i, y_{\text{training}}, x_{\text{training}}) &= \int p(\tilde{y}_i | \tilde{x}_i, \theta) p(\theta | y_{\text{training}}, x_{\text{training}}) d\theta \\ &\approx \sum_{s=1}^{S} p(\tilde{y}_i | \tilde{x}_i, \theta^{(s)}) \end{aligned}$$

where $\theta^{(s)}$ is a sample from the posterior distribution of the parameters ($\theta = \{a, b, \sigma^2\}$ in the original model), that is $\theta^{(s)} \sim p(\theta | y_{\text{training}}, x_{\text{training}})$.

GRADING: Each of the above four tasks provides 5 points from correct implementation and answer. Each task gives 2 points if it is done towards right direction and partially correct.