Bayesian data analysis – exercise 7

The maximum amount of points from this assignment is 6. In addition to the correctness of the answers, the overall quality and clearness of the report is evaluated.

Report all results to a single, **anonymous** *.pdf -file and return it to peergrade.io. Include also source code to the report (either as an attachment or as a part of the answer). By anonymity it is meant that the report should not contain your name or student number.

---

1. Linear model: drowning data with Stan (3p)

The provided `drowning.txt` file contains the number of people drown per year in Finland 1980–2016. We are going to fit a linear model with Gaussian noise to these data using time as the predictor and number of drownings as the target variable (see related linear model example for the Kilpisjärvi-temperature data in the example Stan codes). We have two objective questions:

i) What can you say about the trend in the number of people drown per year? We would plot the histogram of the slope of the linear model.

ii) What does the model predict for year 2019? We would plot the histogram of the posterior predictive distribution for number of people drown at $\tilde{x} = 2019$.

The provided Stan code in Listing 1 is almost correct for the given problem. However, there are two crucial mistakes. Find these two mistakes and fix them. Report the original mistakes and your fixes clearly in your report.

The provided broken code does not define any prior for the parameters. In Stan, this corresponds to using a uniform prior. In addition to the two fixes discussed above, we would like to apply a weakly-informative prior $\mathrm{N}(0, \tau^2)$ for the slope parameter `beta` into the code. It is very unlikely that the mean number of drownings changes more than 50 % in one year. The approximate historical mean yearly number of drownings is 138. Hence, set τ so that the following holds for the prior probability for `beta`: $\Pr(-69 < \texttt{beta} < 69) = 0.99$. Determine suitable value for τ and report the approximate numerical value for it in the report. Using the obtained τ, implement the desired prior in the Stan code. In the report, in a separate section, indicate clearly how you carried out your prior implementation, e.g. "Added line ... in block ...".

Example resulting plots for the problem, with the fixes and the desired prior applied, are shown in Figure 1. If you want, you can use these plots as a reference for testing if your modified Stan code produces similar results. However, running the inference and comparing the plots is not required. In the report, in addition to the full resulting Stan model code, the following things needs to be clearly presented and discussed in separate sections:

- the two fixes in the code,
- suitable approximate numerical value for τ,
- details how to implement the desired prior in the code.

Listing 1: Broken Stan code for question 1

```
data {
    int<lower=0> N;   // number of data points
    vector[N] x;      // observation year
    vector[N] y;      // observation number of drowned
    real xpred;       // prediction year
}
parameters {
    real alpha;
    real beta;
    real sigma;
}
transformed parameters {
    vector[N] mu;
    mu = alpha + beta*x;
}
model {
    y ~ normal(mu, sigma);
}
generated quantities {
    real ypred;
    ypred = normal_rng(mu, sigma);
}
```

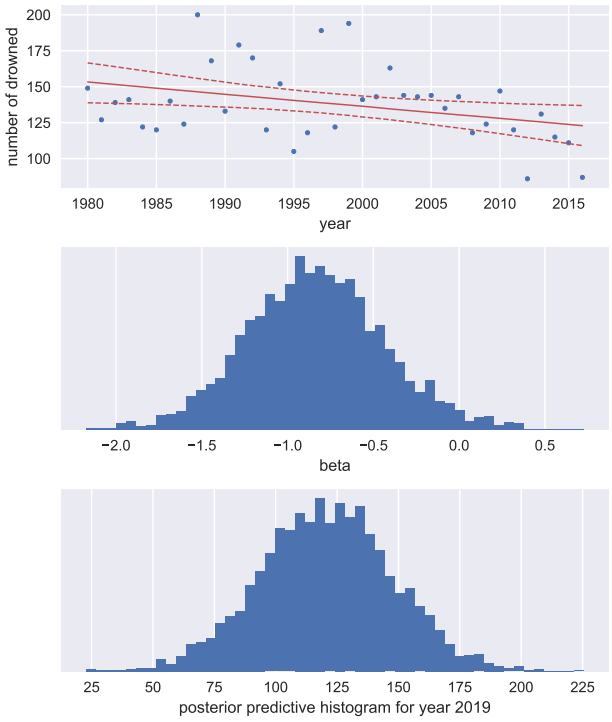

Figure 1: Example plots for the results obtained for problem in the question 1. In the first subplot, the red lines indicate the resulting 5 %, 50 %, and 95 % posterior quantiles for the transformed parameter `mu` at each year.

2. Hierarchical model: factory data with Stan (3p)

The provided `factory.txt` file contains quality control measurements from 6 machines in a factory (units of the measurements are irrelevant here). In the data file, each column contains the measurements for a single machine. Quality control measurements are expensive and time-consuming, so only 5 measurements were done for each machine. In addition to the existing machines, we are interested in the quality of another machine (the seventh machine).

Implement a separate, pooled and hierarchical Gaussian model described in Section 11.6 using Stan. In the pooled model, all the measurements are combined together and no distinction is made between the machines. In the separate model, each machine has its own model. Similarly as in the model description in the book, use the same measurement standard deviation σ for all the groups in the hierarchical model. In the separate model however, use separate measurement standard deviation σ_j for each group j. Use Stan's default uniform prior for all the parameters.

Using each of the three models – separate, pooled, and hierarchical – report (comment and, if applicable, plot histogram):

i) the posterior distribution of the mean of the quality measurements of the sixth machine

ii) the predictive distribution for another quality measurement of the sixth machine

iii) the posterior distribution of the mean of the quality measurements of the seventh machine.

Hint: See the example Stan-codes for the comparison of k groups with and without the hierarchical structure. What you need to do is change the dataset, implement the prediction for the future measurement of the sixth machine, and figure out the distribution for the mean of the quality measurements for the seventh machine in the hierarchical model.