ISYE 6420 – HW #6

Jing Ma

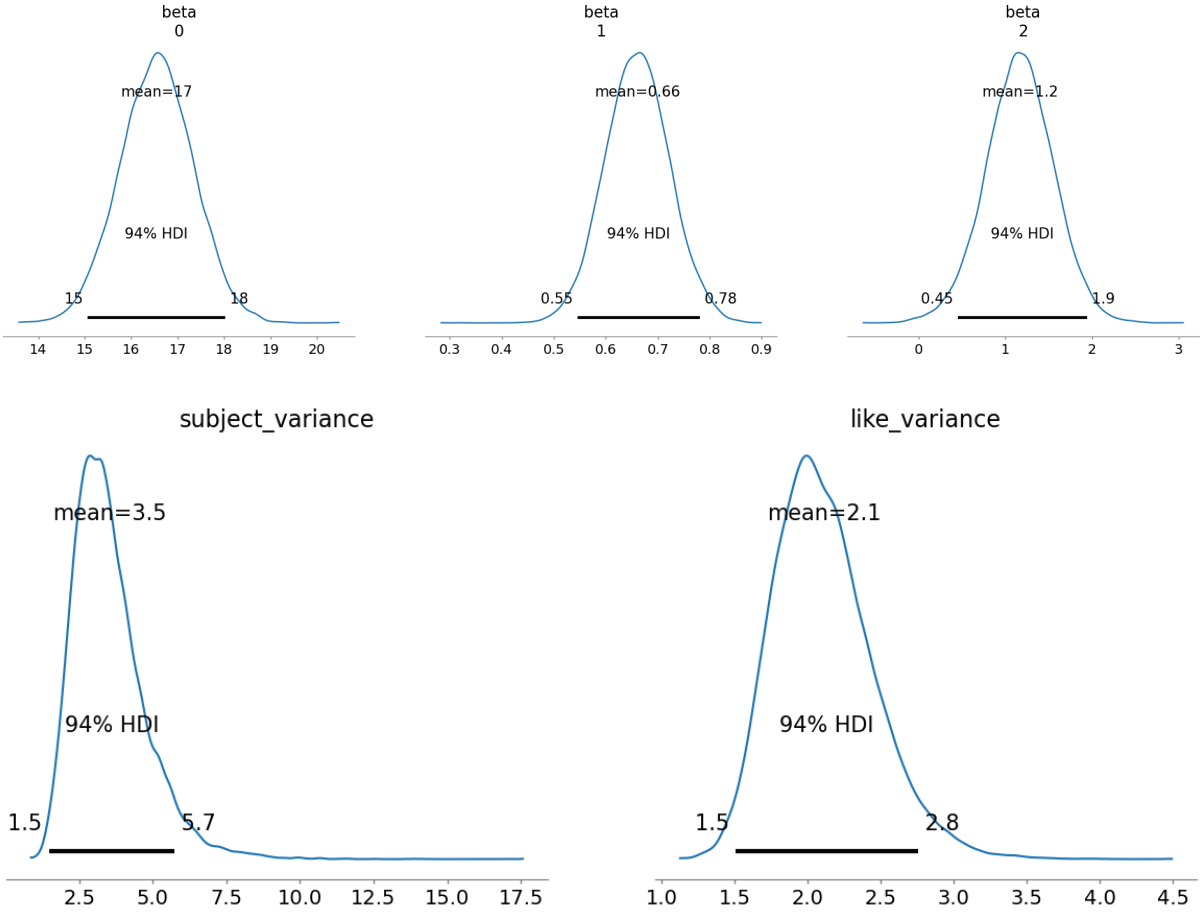
2024-11-16

Problem 1:

1.1

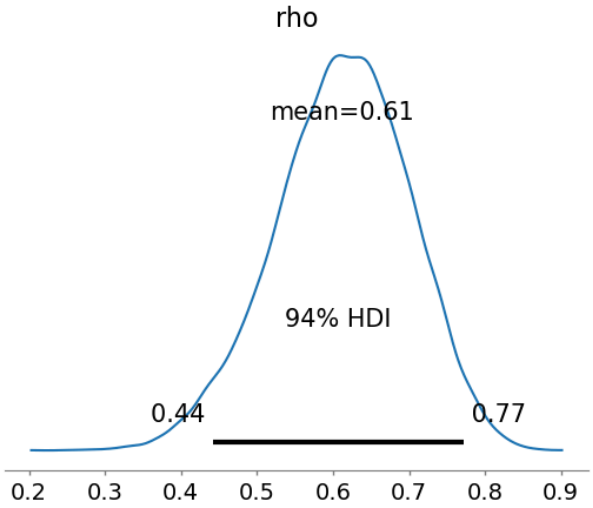
In this problem, we build a linear regression model with random effects included for each subject. The random effects follow the given Normal distribution. Please see the accompanying Jupyter Notebook for the parameters that we used in building the model.

We noted the following posterior density plots for $\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2$, and σ_u^2:



1.2

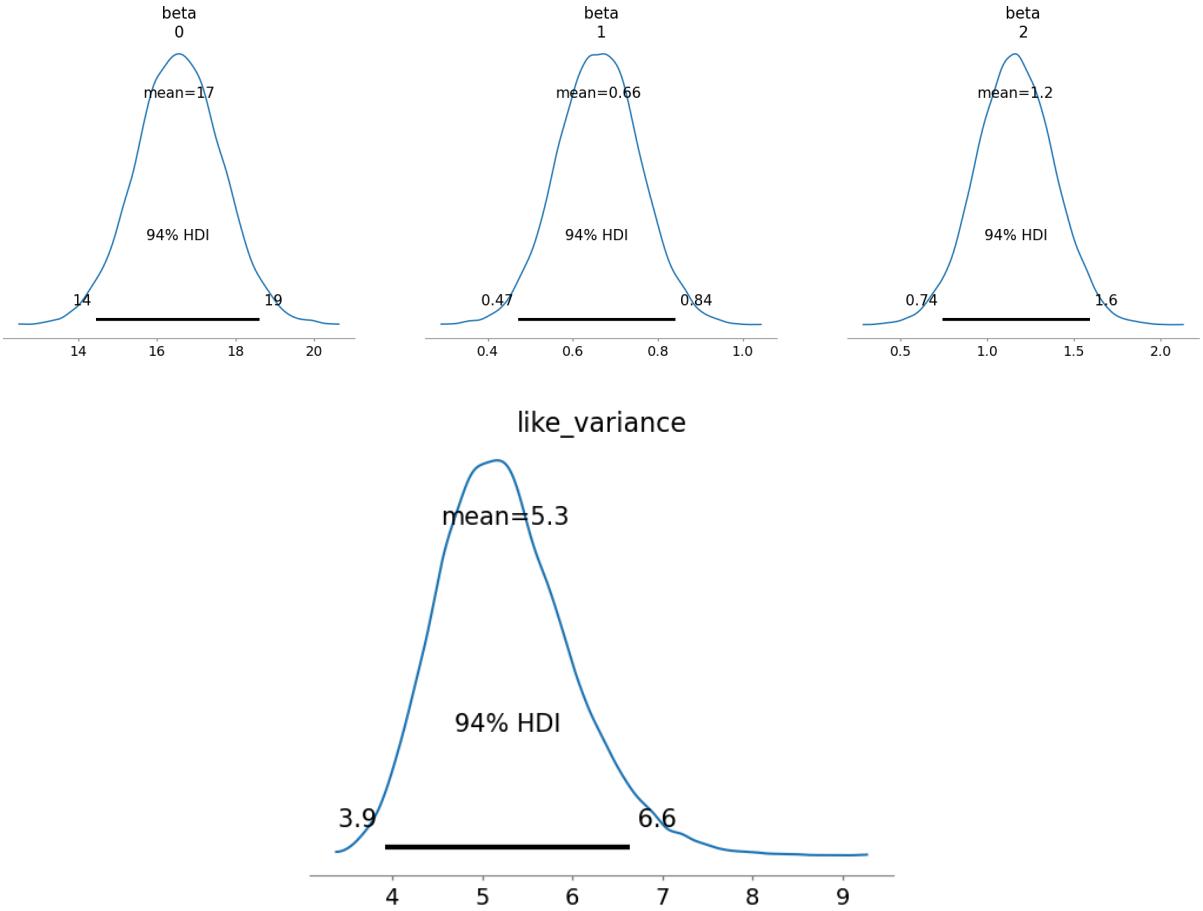
We kept track of ρ, the intraclass correlation coefficient, based on the given equation within the model. Below is the posterior density plot of ρ:



Based on the plot, we can see that the 94% credible set for ρ falls in the interval of [0.44, 0.77]. Since 0 is not within this interval, it suggests that ρ is significantly different from 0 with high credibility.

1.3

In this problem, we simply built the linear regression model without any account for the random effects. Below is the posterior density plots for $\beta_0, \beta_1, \beta_2$, and σ_u^2:



By comparing the plots for the coefficients from part 1 above and herein, we can observe that there is no noticeable difference in the means and 94% credible sets. However, the main difference resides in the mean value and 94% credible sets of the likelihood variance, which increased more than

doubled than with random effects. This result makes sense since previously when we incorporated random effects, some of the variance in the observations would then be proportionally be attributed to σ_u^2. However, since we, in part 2, do not consider the random effects, the variance now has to be explained by likelihood variance, σ_ε^2, in entirety.

Problem 2:

2.1

In Question 2, we are given the species and geographic variables data of the 30 Galapagos islands. In the "Elevation" column, we have some missing data. We performed multiple imputation by using an exponential distribution with $\lambda = \frac{1}{425}$. We also standardized all five variables. Please see the accompanying Jupyter Notebook for the implementation.

We found the following mean and 95% HPD credible intervals of the coefficients:

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta0_intercept	3.887	0.030	3.828	3.943	0.0	0.0	12722.0	13883.0	1.0
beta1_area	-0.494	0.022	-0.536	-0.452	0.0	0.0	12059.0	12889.0	1.0
beta2_elevation	1.463	0.036	1.397	1.532	0.0	0.0	10381.0	11917.0	1.0
beta3_nearest	0.127	0.026	0.077	0.176	0.0	0.0	14183.0	14123.0	1.0
beta4_scruz	-0.424	0.043	-0.504	-0.342	0.0	0.0	13542.0	14032.0	1.0
beta4_adjacent	-0.561	0.025	-0.609	-0.514	0.0	0.0	14776.0	14487.0	1.0

2.2

Based on the summary output above, we can see that elevation is the largest coefficient (not considering the intercept) compared to the other four variables. The other four variables' coefficients appear to suggest moderate to weak effects on the observations, given that their absolute values are below 1.

Problem 3:

3.1

In this problem, we are given the treatment data on 86 patients, who either received placebo or chemotherapy. We are also provided the data on whether cancer recurrence was observed. For the data that has observed value $= 0$, we consider them to be right censored data.

We created the parameters as instructed by the problem and obtained the following output:

	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta0_intercept	-3.282	0.187	-3.671	-2.938	0.002	0.001	13564.0	18354.0	1.0
beta1_treatment	-0.533	0.301	-1.108	0.065	0.003	0.002	13872.0	19089.0	1.0
mu0	27.112	5.208	17.854	37.587	0.046	0.033	13564.0	18354.0	1.0
mu1	46.706	11.593	27.255	69.861	0.070	0.052	29218.0	26413.0	1.0
mu_diff	19.594	12.670	-3.108	45.531	0.097	0.069	16392.0	19541.0	1.0

In the output, the variable, "mu_diff", represents the difference between μ_1 and μ_0. And we can observe that the 95% HPD credible interval of "mu_diff" to be [–3.108, 45.531]. Please see the accompanying Jupyter Notebook for the code implementation.

3.2

To find the posterior probability of the hypothesis $H : \mu_1 > \mu_0$, we are essentially finding the probability that $\mu_1 - \mu_0 > 0$, which in our case, is the probability that $\text{mu\_diff} > 0$.

Therefore, we can find the proportion of mu_diff samples that are greater than 0 and compute it explicitly. Based on our computation, the posterior probability is 0.9647.

3.3

Based on the comparison between μ_1 and μ_0, we can see that 96.47% the time, patients have longer average survival time till cancer recurrence if they received chemotherapy compared to those who received placebo treatment. Further, we observed that the average increase in the survival time is about 19.6 months if a patient received chemotherapy, which is quite significant. In conclusion, the effects of the chemotherapy appear to be very effective on lengthening patients' survival period or can even help combat the potential return of the cancer.