ISYE 6420 – HW #5

Jing Ma

2024-11-03

Problem 1:

1.1

In this question, we are given the PONV dataset where we want to use Gender, Anaesthesiaduration, smoking, and PONVhist to predict the SinclairScore. To create the Bayesian linear regression model, we assumed the following probability distributions and setup. *Please see the accompanying Jupyter Notebook for the implementation of the model.* The code is adapted based on Unit 6.5 Loading Data, Step Function, and Deterministic Variables by Aaron Reding.

$$y_i|\ \mu, \sigma \sim \text{Norm}(\mu_i, \sigma^2)$$

$$\beta_i \sim \text{Norm}(0, 1000^2)$$

$$\tau \sim \text{Gamma}(0.001, 0.001)$$

$$\sigma = \frac{1}{\sqrt{\tau}}$$

$$\mu_i = \beta_0 + \sum_{i=1}^{n=4} \beta_i x_i$$

Based on the code output, we note that the 95% credible set for parameter beta2, the coefficient for Anaesthesiaduration is [0.002, 0.002], which does not include 0. It seems to suggest that we are 95% confident that the true effect of anesthesia duration on the SinclairScore is positive and non-zero and aligns with information provided in the question that the duration of anaesthesia is one of the main risk factors for PONV.

1.2

We made prediction for the single observation as provided in part 2 based on Unit 6.8 Prediction by Aaron Reding. *Please see the accompanying Jupyter Notebook for the prediction.* Per the output, we observed that the 95% credible set for this single observation is [−0.010789, 0.250984].

1.3

We computed Bayesian R^2 based on Unit 6.8 Prediction by Aaron Reding. *Please see the accompanying Jupyter Notebook for the computation of* R^2. Per the result, we observed that the R^2 is 0.485251.

Problem 2:

2.1

In this question, we are given 4 different colors and 6 observed numbers of trapped beetles for each color.

Please see the accompanying Jupyter Notebook for the ANOVA analysis created using PYMC.

The setup of the analysis is as follows:

$$y_i \mid \mu_i, \sigma \sim \text{Norm}(\mu_i, \sigma^2)$$

$$\sigma \sim \text{Inverse-Gamma}(0.1, 0.1)$$

$$\alpha_i \sim \text{Norm}(0, 10^2)$$

$$\mu_0 \sim \text{Norm}(0, 10^2)$$

$$\mu_i = \mu_0 + \alpha_i$$

$$\sum_{i=1}^{n=4} \alpha_i = 0$$

Our null hypothesis states the belief that all treatment groups have the same effect. Therefore, as part of the ANOVA analysis, we compute the differences among all sets of two α_i and examine whether the differences are statistically different from 0.

2.2

	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu0	26.708	1.474	23.823	29.642	0.011	0.008	18712.0	13288.0	1.0
alpha4	3.489	2.455	-1.441	8.309	0.018	0.014	18714.0	14323.0	1.0
alpha3	-11.884	2.406	-16.646	-7.153	0.018	0.013	17986.0	13764.0	1.0
alpha2	-11.055	2.426	-15.761	-6.139	0.018	0.013	18460.0	13579.0	1.0
sigma	7.071	1.195	4.972	9.483	0.010	0.007	14183.0	13180.0	1.0
alpha1	19.451	2.502	14.602	24.450	0.015	0.011	26946.0	15837.0	1.0
alpha1 - alpha2	30.506	4.046	22.423	38.272	0.026	0.019	23456.0	17364.0	1.0
alpha1 - alpha3	31.335	4.020	23.529	39.467	0.026	0.019	23252.0	15813.0	1.0
alpha1 - alpha4	15.962	4.058	7.668	23.779	0.026	0.019	23693.0	17040.0	1.0
alpha2 - alpha3	0.829	3.903	-6.818	8.492	0.030	0.029	17105.0	13181.0	1.0
alpha2 - alpha4	-14.544	3.988	-22.256	-6.543	0.030	0.022	17663.0	13108.0	1.0
alpha3 - alpha4	-15.373	3.964	-23.186	-7.495	0.029	0.021	18497.0	13081.0	1.0

Based on the output above from the Notebook, we noticed the following:

1. Treatment Group “Lemon Yellow” has higher mean coefficient than the other three groups (“White”, “Green”, “Blue”), which suggests that “Lemon Yellow” attracts the most beetles.
2. Treatment Group “White” has a mean coefficient that is marginally higher than “Green”, which seems to suggest that the treatment effects between these two groups are quite close.
3. Treatment Group “White” has lower mean coefficient compared to “Blue”, which appears that “Blue” attracts more beetles than “White”.
4. Similarly, “Green” group shows less attraction to beetles compared to “Blue” given the lower mean coefficient.

Problem 3:

3.1

We performed three methods to fit a frequentist logistic regression given the Iris dataset. *For more detailed implementation, please see the accompanying Jupyter Notebook.*

1. We fit the logistic regression without any regularization using statsmodels package
2. We fit the logistic regression with Lasso regularization (L1) using statsmodels package

3. We fit the logistic regression with Ridge regularization (L2) using sklearn package combined with bootstrapping method

3.1.a

First, let's take a look at the frequentist logistic regression without any regularization. We noticed in the output below that each independent variables as well as the intercept has an extremely wide 95% credit set. This suggests that there is very high uncertainty in the parameter estimates. One of the possible causes is multicollinearity, where the predictors maybe highly correlated with each other and hence adding uncertainty to the coefficients' possible range.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	9.6813	1.2e+04	0.001	0.999	-2.35e+04	2.35e+04
Sepal_Length	-4.1173	3316.583	-0.001	0.999	-6504.500	6496.265
Sepal_Width	-8.9814	1815.027	-0.005	0.996	-3566.370	3548.407
Petal_Length	4.4103	1631.257	0.003	0.998	-3192.795	3201.615
Petal_Width	33.8138	5245.781	0.006	0.995	-1.02e+04	1.03e+04

3.1.b

Now let's check the summary output when we apply L1 regularization. As we can see below, Lasso didn't help with reducing the credible set ranges for the coefficients. This might be that Lasso is typically used for variable selection rather than reducing uncertainty and that multicollinearity is not being adequately addressed.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.9979	9.39e+06	-2.13e-07	1.000	-1.84e+07	1.84e+07
Sepal_Length	-8.2904	3.78e+06	-2.19e-06	1.000	-7.42e+06	7.42e+06
Sepal_Width	-14.3042	7.15e+06	-2e-06	1.000	-1.4e+07	1.4e+07
Petal_Length	28.5257	2.95e+06	9.68e-06	1.000	-5.77e+06	5.77e+06
Petal_Width	11.7497	7.26e+06	1.62e-06	1.000	-1.42e+07	1.42e+07

3.1.c

Next, we will take a look at the summary output from using Ridge regularization. As we can see, the bounds of 95% credible sets have been significantly reduced, showing more numerical stability in the model fitting process.

	Feature	Coefficient	Lower CI	Upper CI
0	Intercept	-0.258227	-0.007384	0.005562
1	Sepal_Length	-0.402121	0.353185	0.581671
2	Sepal_Width	-1.464355	-1.018575	-0.684319
3	Petal_Length	2.237124	2.243529	2.353407
4	Petal_Width	1.000677	0.855301	1.054218

3.2

Now we will use Bayesian logistic regression method with an uninformative prior $N(0, 1000)$. Though the 94% credible set ranges are not as wide as the frequentist method, it's still quite noticeably large.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
alpha	-4.067	30.271	-60.450	53.244	0.576	0.408	2761.0	3851.0	1.0
betas[0]	-6.796	16.005	-38.372	21.725	0.388	0.274	1682.0	2877.0	1.0
betas[1]	-22.193	21.663	-62.381	18.886	0.492	0.348	1959.0	3132.0	1.0
betas[2]	36.006	19.942	2.565	75.750	0.464	0.328	1831.0	3538.0	1.0
betas[3]	16.860	27.322	-34.338	67.930	0.500	0.353	3001.0	4349.0	1.0

3.3

When we re-ran the Bayesian logistic regression model using an informative prior $N(0, 1)$, we notice much smaller 94% credible set ranges which is very similar to the effect that we observed from using L2 regularization above.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
alpha	-0.274	0.981	-2.070	1.637	0.009	0.007	11214.0	10744.0	1.0
betas[0]	-0.430	0.587	-1.546	0.654	0.007	0.005	7804.0	9641.0	1.0
betas[1]	-1.585	0.747	-2.964	-0.142	0.008	0.006	8272.0	9456.0	1.0
betas[2]	2.423	0.624	1.238	3.570	0.007	0.005	8581.0	9255.0	1.0
betas[3]	1.100	0.927	-0.593	2.880	0.009	0.007	10338.0	10294.0	1.0

Therefore, by comparing the results above, we noticed that Bayesian LR model using the informative prior produces the most meaningful results. This is because $N(0, 1)$ effectively regularizes the coefficient estimates from becoming too large by pulling them toward zero. In contrast, $N(0, 1000)$ as a vague prior provides little to no useful information to the model.