MBA5011 Multivariate Analysis: Model-Based Statistics

Assignment 4

Question 1 (50%).

Consider the `data(rugged)` data (in `rethinking` package) on economic development and terrain ruggedness examined in Chapter 7. One of the African countries in the example, Seychelles, is far outside the cloud of other nations, being a rare country with both relatively high GDP and high ruggedness. Seychelles is also unusual, in that it is a group of islands far from the coast of mainland Africa, and its main economic activity is tourism. One might suspect that this one nation is exerting a strong influence on the conclusions. In this problem, we want you to drop Seychelles from the data and re-evaluate the hypothesis that the relationship of African economies with ruggedness is different from that on other continents.

(a) Begin by using `rstan` to fit just the interaction model(use `print` to show your coefficient):
$y_i \sim Normal(\mu_i, \sigma)$
$\mu_i = \alpha + \beta_A A_i + \beta_R R_i + \beta_{AR} A_i R_i$

where y is log GDP per capita om the year 2000 (log of `rgdppc_2000`); A is cont_africa, the dummy variable for being an African nation; and R is the variable rugged. Choose your own priors. Compare the inference from this model fit to the data without Seychelles to the same model fit to the full data. Does it still seem like the effect of ruggedness depends upon continents? How much has the expected relationship changed?

(b) Now plot the predictions of the interaction model, with and without Seychelles. Does it still seem like the effect of ruggedness depends upon continents? How much has the expected relationship changed?

(c) Finally, conduct a model comparison analysis, using WAIC. Fit three models to the data without Seychelles.

Model1:

$y_i \sim Normal(\mu_i, \sigma)$

$\mu_i = \alpha + \beta_R R_i$

Model2:

$y_i \sim Normal(\mu_i, \sigma)$

$\mu_i = \alpha + \beta_A A_i + \beta_R R_i$

Model3:

$y_i \sim Normal(\mu_i, \sigma)$

$\mu_i = \alpha + \beta_A A_i + \beta_R R_i + \beta_{AR} A_i R_i$

Use whatever priors you think are sensible. Plot the model-averaged predictions of this model set. Do your inferences differ from those in (b)? Why or why not?

Question 2 (50%).

value in data(nettle) are data on language diversity in 74 nations. The meaning of each column is given below:

(1) `country`: Name of the country
(2) `num.lang`: Number of recognized language spoken
(3) `area`: Area in square kilometers
(4) `k.pop`: Population in thousands
(5) `num.stations`: Number of weather stations that provided data for the next two columns
(6) `mean.growing.season`: Average length of growing season, in months.
(7) `sd.growing.season`: Standard deviation of length of growing season, in months.

Use these data to evaluate the hypothesis that language diversity is partly a product of food security. The notion is that, in productive ecologies, people don't need large social networks to buffer them against risk of food shortfalls. This means ethnic groups can be smaller and more self-sufficient, leading to more languages per capita. In contrast, in a poor ecology, there is more subsistence risk, and so human societies have adapted by building larger networks of mutual obligation to provide food insurance. This in turn creates social forces that help prevent languages from diversifying.

Specifically, you will try to model the number of languages per capita as the outcome variable:

```
d<-
  d %>%
    mutate(lang.per.cap = num.lang/k.pop)
```

Using the logarithm of this new variable as your regression outcome. This problem is open ended, allowing you to decide how you address the hypotheses and the uncertain advice the modeling provides. If you think you need to use WAIC anyplace, please do. If you think you need certain priors, argue for them. If you think you need to plot predictions in a certain way, please do. Just try to honestly evaluate the main effects of both `mean.growing.season` and `sd.growing.season`, as well as their two-way interaction, as outlined in parts (a), (b) and (c) below.

(a) Evaluate the hypothesis that language diversity, as measured by `log(lang.per.cap)`, is positively associated with the average length of the growing season, `mean.growing.season`. Consider `log(area)` in your regression(s) as a covariate (not an interaction). Interpret your results.

(b) Now evaluate the hypothesis that language diversity is negatively associated with the standard deviation of length of growing season, `sd.growing.season`. This hypothesis follows from uncertainty in harvest favoring social insurance through larger social networks and therefore fewer languages. Again, consider `log(area)` as a covariate (not an interaction). Interpret your results.

(c) Finally, evaluate the hypothesis that `mean.growing.season` and `sd.growing.season` interact to synergistically reduce language diversity. The idea is that, in nations with longer average growing seasons, high variance makes storage and redistribution even more important than it would be otherwise. That way, people can cooperate to preserve and protect windfalls to be used during the droughts. These forces in turn may lead to greater social integration and fewer languages.