Model selection and multimodel inference

QERM 514 - Homework 5

1 May 2020

Background

This week's home work will require you to use all of the information you have learned so far in class. Your task is to analyze some data on the concentration of nitrogen in the soil at 41 locations on the island of Maui in the Hawaiian Archipelago. Along with the nitrogen measurements, there are 4 possible predictor variables that may help to explain the variation in soil nitrogen concentration. The accompanying data file `soil_nitrogen.csv` has the following 5 columns of data:

- `nitrogen`: concentration of soil nitrogen (mg nitrogen kg^{-1} soil)
- `temp`: average air temperature (°C)
- `precip`: average precipitation (cm)
- `slope`: slope of the hillside (degrees)
- `aspect`: aspect of the hillside (N, S)

As you work through the following problems, be sure to show all of the code necessary to produce your answers.

Problems

a) Begin by building a global model that contains all four of the predictors plus an intercept. Show the resulting ANOVA table, and report the multiple and adjusted R^3 values. Also report the estimate of the residual variance $\hat{\sigma}^2$.

b) Check the residuals from your full model for possible violations of the assumption that the $e_i \sim \mathrm{N}(0, \sigma^2)$.

c) Does this seem like a reasonable model for these data? Why or why not?

d) Now fit various models using all possible combinations of the 4 predictors, including an intercept-only model (ie, there should be a total of 16 models). Compute the AIC, AICc, and BIC for each of your models and compare the relative rankings of the different models.

e) Conduct a leave-one-out cross-validation for all of the models in part (d), using the root mean squared prediction error (RMSPE) as your scale-dependent measure of fit. Report your results alongside your results from part (d). Do all of the methods agree on which of these models is the best?

f) Given some uncertainty that one of these models is the true data-generating model, compute the weights of evidence for each of the models in your set. Which model has the greatest support from the data? What are the odds against the intercept-only model compared to the best model?

g) Calculate the model-averaged parameters across all models in your set. Use these parameters to predict what the soil nitrogen concentration would be on the nearby island of Moloka'i if the average precipitation was 150 cm, the average temperature was 22 °C, and the hillside faced south with a slope of 11 degrees.

h) Compare your prediction from part (g) to a prediction from the model identified as the best in part (e), using the same inputs. How much do they differ from one another?