ISYE 6420 (Bayesian Statistics) Project: apremkumar7

Bayesian Linear Regression to Predict Residuary Resistance of Ships

Introduction

Prediction of residuary resistance of yachts and ships at the time of design and construction is of incredible value to estimate the ship's resistance and evaluate how much propulsive power is required. Residuary resistance is the resistance opposing the motion of a ship through the water and what remains after frictional resistance is subtracted from fluid resistance.

Linear regression can be used to predict this value using hull dimensions and velocity as predictors. This project is an attempt at using JAGS (Just Another Gibbs Sampler) in R to implement bayesian linear regression, compare various models and predict future values using the model with best performance indicators.

Dataset

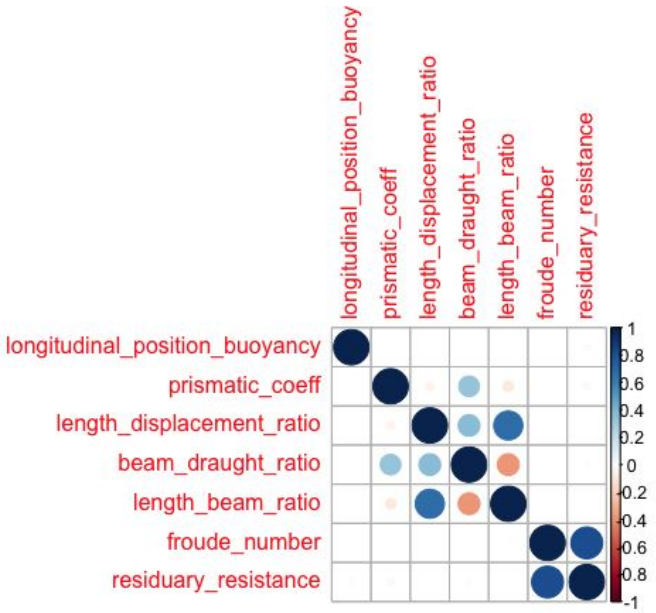
A study was conducted at the Delft Ship Hydromechanics Laboratory at Technical University of Delft to examine this problem that includes 308 experiments [1]. The dataset is available online via UC Irvine Machine Learning Repository [2] . There are 6 variables provided as predictors in this dataset, most of them being hull dimensions, since hull is the main body of the ship. They are: 1. Longitudinal position of the center of buoyancy, 2. Prismatic coefficient, 3. Length-displacement ratio, 4. Beam-draught ratio, 5. Length-beam ratio, and 6. Froude number. The response variable is Residuary resistance per unit weight of displacement.

Bayesian Modeling

JAGS in R is used to implement bayesian inference of this model. The entire code is attached to this document. The following sections outline the steps followed to model this problem and arrive at conclusions.

1. Dataset Summary/Correlations:

Summary statistics are examined using summary function in R. The correlation matrix chart is plotted to investigate any potential auto-correlation problems in linear regression.



In this chart, the bigger the circle the higher the correlation. It is noticed that froude number has the highest correlation with residuary resistance, hence with highest predictive power. Length displacement ratio and length beam are relatively highly and positively correlated. In subsequent steps, length beam ratio is excluded from the list of predictors to build the model in an effort to possibly improve metrics. Dataset is also split into training and test to perform cross validation at a later stage, with 80% assigned to train and 20% to test.

2. JAGS Modeling:

In an effort to try out various models using the training dataset and compare them using model performance indicators to find the best one, 3 scenarios are identified and pursued as part of modeling. All of them follow the same approach. JAGS model is defined by specifying a likelihood where the response variable is a normal random variable, with mean equal to the linear combination variables with coefficients. Priors are then provided for each of the coefficients, all of them being non-informative at the moment. Parameters are

defined and initialized. The JAGS model is then fit to the data. The first 2000 iterations are burnt and 10000 are captured. Deviance Information Criterion (DIC) is used to select the best model.

a. With all variables

The first model includes all variables, irrespective of correlation values. Deviance Information Criterion (DIC): 1880.2

b. With length beam ratio variable removed due to possible autocorrelation

One of the requirements for linear regression to work is for autocorrelation to not exist between variables. For that reason, using the correlation plot mentioned above, length beam ratio is identified as one with relatively high correlation with length displacement ratio and is removed from this model. DIC turns out to be: 1875.5

c. With informative priors

The third model is one with informative priors (normal distribution with actual mean and variance calculated). But it turns out to be the worst of all, with DIC: 2046

3. Model Comparison

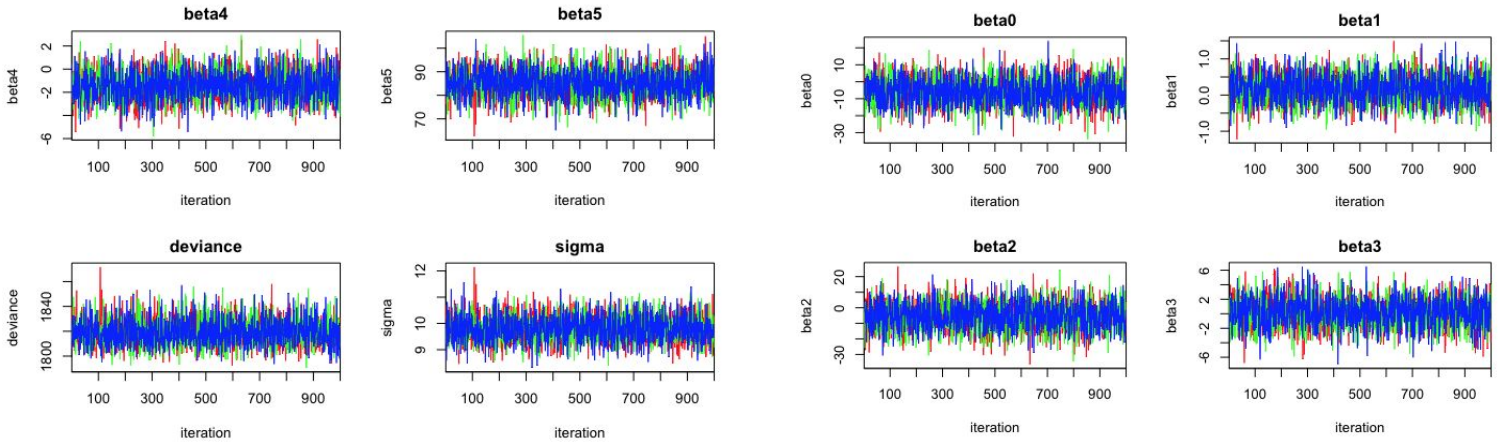
To choose the optimal model, DIC is used to compare between the three and the lowest one (one with length beam ratio excluded) is chosen. Here are the estimators for various means of coefficients, intercept, standard dev and deviance, with 95% credible intervals:

```
Inference for Bugs model at
"/var/folders/w2/0h_967js7fld5dnx4twmv8wh0000gn/T//RtmpL7axDV/model1029e7849361c.txt", fit using
jags,
 3 chains, each with 12000 iterations (first 2000 discarded), n.thin = 10
 n.sims = 3000 iterations saved
          mu.vect sd.vect     2.5%      25%      50%      75%    97.5%  Rhat n.eff
beta0      -5.845   7.986  -20.974  -11.081   -6.015   -0.584    9.797 1.001  3000
beta1       0.176   0.402   -0.599   -0.104    0.178    0.442    0.953 1.001  3000
beta2      -5.496   8.772  -22.634  -11.227   -5.463    0.519   11.489 1.001  3000
beta3       0.154   2.000   -3.728   -1.171    0.202    1.545    4.037 1.001  2600
beta4      -1.417   1.270   -3.838   -2.256   -1.445   -0.545    1.069 1.001  2100
beta5      85.198   5.801   73.534   81.426   85.191   89.037   96.294 1.001  3000
sigma       9.800   0.496    8.866    9.451    9.790   10.131   10.808 1.001  3000
deviance 1819.170  10.614 1800.150 1811.697 1818.256 1825.805 1842.161 1.001  3000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 56.4 and DIC = 1875.5
DIC is an estimate of expected predictive error (lower deviance is better).
```

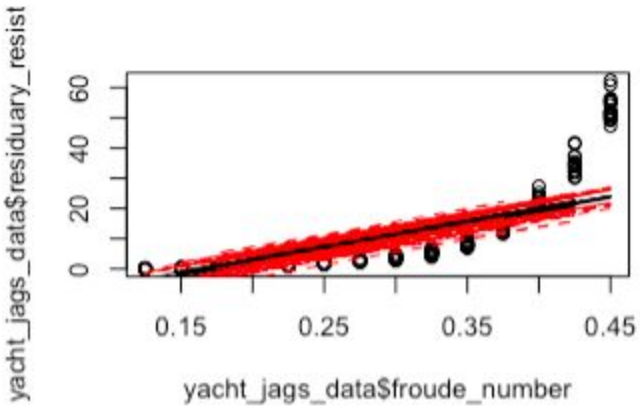
From the above table, one important observation is the mean of beta5 (coefficient of Froude Number). A value of 85.19 means an unit increase of Froude Number results in a mean increase of Residuary Resistance, keeping all other variables constant.

4. Plot Results

The results of the fitted model are plotted below. To make sure the MCMC chains are mixed properly, traceplot function is used. From the charts below, it can be seen that the chains are all well mixed.



Cross-validation (similar to a frequentist approach) is performed to test the model chosen using future (20% new observations in test dataset) data. The predicted values and the input (training) dataset in one chart . As part of this, the 3 MCMC chains are merged and new mean Residuary Resistance values are estimated using the posterior values of the parameters. The upper and lower credible intervals wrt to the mean predicted values are also estimated and plotted in red.



As an example, Froude Number is used to plot (as it has the highest correlation with response variable).

References

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2. J. Gerritsma, R. Onnink, and A. Versluis. Geometry, resistance and stability of the delft systematic yacht hull series. In International Shipbuilding Progress, volume 28, pages 276â€“297, 1981.
3. I. Ortigosa, R. Lopez and J. Garcia. A neural networks approach to residuary resistance of sailing yachts prediction. In Proceedings of the International Conference on Marine Engineering MARINE 2007, 2007.