MATH 4432 Mini-Project 1: Linear Regression Models on Animal Species Sleeping Hours

Tong Chun Ho, Lai Cheuk Man and Wong Ngo Cheung {chtongaa, cmlaiad, ncwongad}@ust.hk
Department of Mathematics, HKUST

1. Introduction

We filled in the missing data by median and built 4 generalised linear regression model. Then we estimated the test error by LOOCV and chose the best model. Finally, we use bootstrap for quantification of uncertainty in the model.

2. Sleep Dataset

Method:

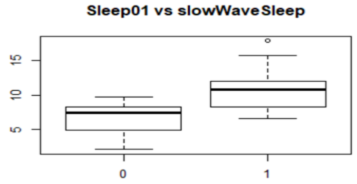
- ➢ Filling in the missing values by median

Reason:.

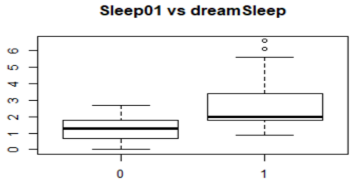
- ➢ As the number of data is limited. Removing the missing data will cause a prediction model to easily pick up the patterns caused by random chance.

Observation:

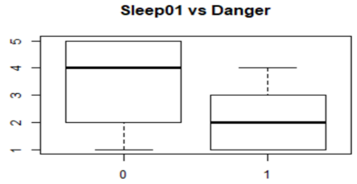
- ➢ ‘slowWaveSleep’, ‘dreamSleep’, ‘danger’ and ‘predation’ may have relationships with sleep01.



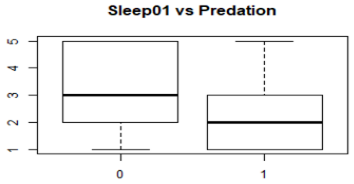
Sleep01 vs slowWaveSleep



Sleep01 vs dreamSleep



Sleep01 vs Danger



Sleep01 vs Predation

3. Model

Method:

- ➢ Generalised linear regression

Reason:

- ➢ The data contain categorical factors and quantitative factors.

Model:

1) slowWaveSleep + dreamSleep + danger + predation
2) slowWaveSleep + dreamSleep + sq_danger + sq_pred (transformation)
3) slowWaveSleep + dreamSleep + pred_danger (interacting)
4) slowWaveSleep + dreamSleep + danger (excluding predation)

Conclusion:

1) danger and predation have very high p-values.
2) The transformation terms have high p-values.
3) The interacting term has a high p-values.
4) The p-values of the predictors are within 0.05.

4. Cross Validation

Method:

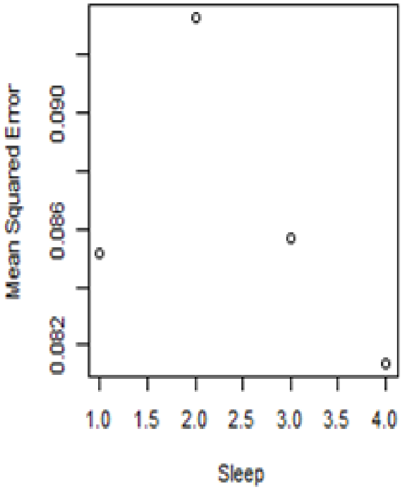
- ➢ Leave-One-Out Cross Validation (LOOCV)

Reason:

- ➢ A small sample size does not need much computing power for the computation of LOOCV. The test error can contain less bias.

Observation:

- ➢ We can discover that the forth model has the smallest test error among all.



5. Bootstrap

Using bootstrap method, we can find out that the standard error is very large.

	Intercept	slowWave Sleep	dream-Sleep	danger
Standard Error	1268.2208	134.5684	101.7052	23.4019

6. Conclusion

Using generalized linear model, we find that in sleep dataset, ‘slowWaveSleep’, ‘dreamSleep’ and ‘danger’ have relationships with ‘sleep’. By using LOOCV, the test error is small. However, the bootstrap statistics show the large standard error.

7. References

James, Witten, Hastie and Tibshirani, "An Introduction to Statistical Learning, with applications in R." (2017).

8. Contribution

Model

- ➢ Tong Chun Ho

Cross Validation

- ➢ Lai Cheuk Man

Bootstrap

- ➢ Wong Ngo Cheung