MATH4432 project1

***Relationship***

***of Animal Species Sleeping Hours***

***with related biological features***

WONG,Chung Wai (20197330)

WONG,Chun Kit(20267587)

**Contents**

1. **Introduction** P.3
2. **Data Analysis** P.3
3. **Cross Validation** P.5

1. **Bootstrap** P.5
2. **Appendix** P.7
3. ***Introduction***

In this report, we are going to discuss the relationship between the sleeping hours. In the biological aspect, animals’ behaviors are particular depends on the gene, for example, dogs always urinate along the street; animals which know to use their hands(using machine) are more clever, and we are hoping to find some relationship between the sleeping hours of animals and their features.

1. ***Data analyst***

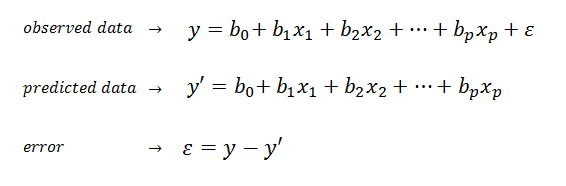
**1.Handling of omission of the data**

We have a set of 62 different species data, including their average slow wave sleeping hours, dream sleeping hours, sleeping hours, body size, brain size, life length, gestation period, predation, sleep exposure and the corresponding danger level. However, some of the data is omitted, therefore, we only use the completed data which has 51 different species (except the omission in wave sleeping hours and dream sleeping hours). We choose to delete the data instead of completing the date because we only delete 11 data which doesn’t take a large proportion in the whole data so it should not have a big impact on our conclusion.

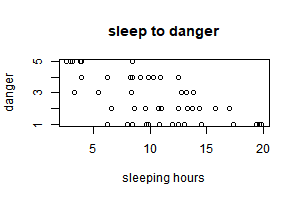
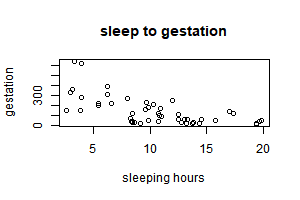
***2.choice of modelling***

In our attempts, we tried to plot the sleeping hours of the animals associating

with their features. Although we can speculate some relationship(see figure2.1), however, most of them show discrete features and it shows that there is no simple significant relationship between sleeping hours and the features (figure 2.2). Therefore, multiple linear regression is more appropriate for this data. Although multiple linear regression is a model with low flexibility, simpler model would provide a much clear concept on our topic. Therefore, we use the following model to deal with our topic:



***Figure 2.1 Figure2.2***



Since there are two types of data ( real-valued and categorical) in data set, transformation of categorical data to dummy variable is essential for multivariable linear regression analysis.

After transforming into a set of new data, estimation of the sleeping hours() are shown in the below:

,

where , , and is the corresponding coefficient,

is the real-valued variables, is body weight, is brain size, is life, is gestation, is the index of dummy variable of predation, where =1 if the specie belong to predation class ,=0 otherwise.

is the index of dummy variable of sleep exposure, where =1 if the specie belong to sleep exposure class ,=0 otherwise.

is the index of dummy variable of danger, where =1 if the specie belong to danger class ,=0 otherwise.

By using R to estimate the model, we get

By simple calculation, we can find the sleeping hours for the animals with class 5 in both predation, sleep exposure and danger is.

In conclusion, we can expect the animals have a longer sleeping hours with lighter body weight, a bigger brain, shorter life and shorter gestation period.

Besides, animal with higher predation class, lower sleep exposure level and higher danger level usually sleep longer.

***3.Testing MSE***

By using this model, we get the MSE=148187, which is very large and which indicates that our model is not so fit to the data, however, it may due to the classification of the data.

***4.Bootstrap***

By working the bootstrap, our model have small bias and standard error for , which is the coefficient of real-valued data, it means the model are accuracy to estimate the , however, because of the dummy variables are discrete, the coefficient of categorical variables are with high bias and standard error.

***Appendix: code for R***

###Amended sleep1 data set discrete variables into dummy variables in another program

> sleep = read.csv("C:/Users/ChunKit/Documents/temp.csv",quote = "")

### doing MLR

> sleep <- sleep[-c(4,13,19,20,21,31,35,36,41,56,62),-c(1,2,3)]

> lm.fit = lm(sleep~.,data=sleep)

> summary(lm.fit)

Call:

lm(formula = sleep ~ ., data = sleep)

Residuals:

Min 1Q Median 3Q Max

-5.2801 -1.3877 -0.2213 1.0472 6.6636

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.838e+00 1.713e+00 4.575 6.07e-05 \*\*\*

body -4.376e-04 1.632e-03 -0.268 0.79024

brain 1.023e-03 1.904e-03 0.537 0.59464

life -1.495e-02 4.061e-02 -0.368 0.71502

gestation -1.702e-02 6.953e-03 -2.449 0.01966 \*

is\_predation\_1 -1.008e+01 4.032e+00 -2.501 0.01737 \*

is\_predation\_2 -4.902e+00 3.382e+00 -1.450 0.15634

is\_predation\_3 -2.626e+00 2.518e+00 -1.043 0.30442

is\_predation\_4 3.790e-01 1.941e+00 0.195 0.84636

is\_sleepExposure\_1 -1.863e+00 4.176e+00 -0.446 0.65827

is\_sleepExposure\_2 -2.426e+00 3.974e+00 -0.610 0.54563

is\_sleepExposure\_3 -2.779e+00 3.564e+00 -0.780 0.44095

is\_sleepExposure\_4 -5.187e-01 3.317e+00 -0.156 0.87665

is\_danger\_1 1.856e+01 5.670e+00 3.273 0.00245 \*\*

is\_danger\_2 1.261e+01 5.138e+00 2.453 0.01945 \*

is\_danger\_3 7.179e+00 4.664e+00 1.539 0.13300

is\_danger\_4 5.647e+00 3.529e+00 1.600 0.11878

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.78 on 34 degrees of freedom

Multiple R-squared: 0.7599, Adjusted R-squared: 0.647

F-statistic: 6.727 on 16 and 34 DF, p-value: 1.68e-06

###croos validation

> set.seed(1)

> train=sample(51,26)

> lmtrain.fit=lm(sleep~.,data=sleep,subset=train)

> mean((sleep-predict(lm.train,sleep))[-train]^2)

[1] 148187.9

> mean((sleep-predict(lm.train,sleep))[-train]^3)

[1] 862643765

###bootstrip

> library(boot)

> boot.fn=function(data,index)

+ return(coef(lm(sleep~.,data=data,subset=index)))

> boot.fn(sleep,1:51)

(Intercept) body brain life

7.837506e+00 -4.376106e-04 1.022973e-03 -1.495358e-02

gestation is\_predation\_1 is\_predation\_2 is\_predation\_3

-1.702459e-02 -1.008407e+01 -4.902114e+00 -2.625696e+00

is\_predation\_4 is\_sleepExposure\_1 is\_sleepExposure\_2 is\_sleepExposure\_3

3.789859e-01 -1.863479e+00 -2.425989e+00 -2.778734e+00

is\_sleepExposure\_4 is\_danger\_1 is\_danger\_2 is\_danger\_3

-5.186613e-01 1.855634e+01 1.260530e+01 7.178776e+00

is\_danger\_4

5.647091e+00

> set.seed(1)

> boot.fn(sleep,sample(51,51,replace=T))

(Intercept) body brain life

9.357649714 -0.002634275 0.003307433 -0.068151907

gestation is\_predation\_1 is\_predation\_2 is\_predation\_3

-0.014126426 -13.463361175 -7.326797898 -5.628424878

is\_predation\_4 is\_sleepExposure\_1 is\_sleepExposure\_2 is\_sleepExposure\_3

-1.085411851 -4.508930384 -4.946701208 -5.615493922

is\_sleepExposure\_4 is\_danger\_1 is\_danger\_2 is\_danger\_3

-0.329346432 22.324310936 16.742288104 9.842202893

is\_danger\_4

5.340193249

> boot(sleep,boot.fn,1000)

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

boot(data = sleep, statistic = boot.fn, R = 1000)

Bootstrap Statistics :

original bias std. error

t1\* 7.837506e+00 -1.679491151 2.437113036

t2\* -4.376106e-04 0.004896010 0.008446346

t3\* 1.022973e-03 -0.008624953 0.012396804

t4\* -1.495358e-02 0.100298948 0.154 903831

t5\* -1.702459e-02 0.001290856 0.010108284

t6\* -1.008407e+01 3.058794597 18.596238511

t7\* -4.902114e+00 2.586064463 16.362122869

t8\* -2.625696e+00 0.491763529 10.634199698

t9\* 3.789859e-01 -0.510125150 7.742144178

t10\* -1.863479e+00 -11.039184292 31.130116683

t11\* -2.425989e+00 -11.400616631 31.491690398

t12\* -2.778734e+00 -11.844725932 31.492200337

t13\* -5.186613e-01 -11.430609544 30.461898622

t14\* 1.855634e+01 9.525402579 34.142354875

t15\* 1.260530e+01 12.819177969 35.343388967

t16\* 7.178776e+00 12.647831981 34.286199061

t17\* 5.647091e+00 13.324667485 31.714912887