STAT_6021_Project2_Proposal

Part 2: Proposal

Each group will submit a project proposal for approval by the instructor, to ensure that an appropriate data set is being used and an appropriate amount of analysis will be done to produce a successful 4 week project. Each proposal should provide:

- The associated data set for the project should be provided, in one of the following ways:

1) providing the names of the R package and the R dataframe;

R Package: "MASS"

R dataframe: "crabs"

```
# from MASS package
`?`(crabs)
head(crabs)
```

```
##   sp sex index   FL  RW   CL   CW  BD
## 1  B   M     1  8.1 6.7 16.1 19.0 7.0
## 2  B   M     2  8.8 7.7 18.1 20.8 7.4
## 3  B   M     3  9.2 7.8 19.0 22.4 7.7
## 4  B   M     4  9.6 7.9 20.1 23.1 8.2
## 5  B   M     5  9.8 8.0 20.3 23.0 8.2
## 6  B   M     6 10.8 9.0 23.0 26.5 9.8
```

2) a link to the data set;

Data can be found at: url(https://r-data.pmagunia.com/dataset/r-dataset-package-mass-crabs)

3) a file containing the data, as an Excel spreadsheet, a .csv file, or a .txt file. Please note that a list of some publicly available sources of data sets is provided on Collab.

```
write.csv(crabs, "crab.data.csv")
```

Image adapted from: Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology 22, 417–425.

• Project objectives/goals. What questions is the group trying to answer, as well as potential practical implications (the more interesting and/or practical, the better) of the results. Your project should involve both linear regression and logistic regression, so clearly state the response variables involved.

• Some data visualizations and commentary related to the project objectives/goals. At least one visualization related to linear regression, and at least one visualization related to logistic regression.

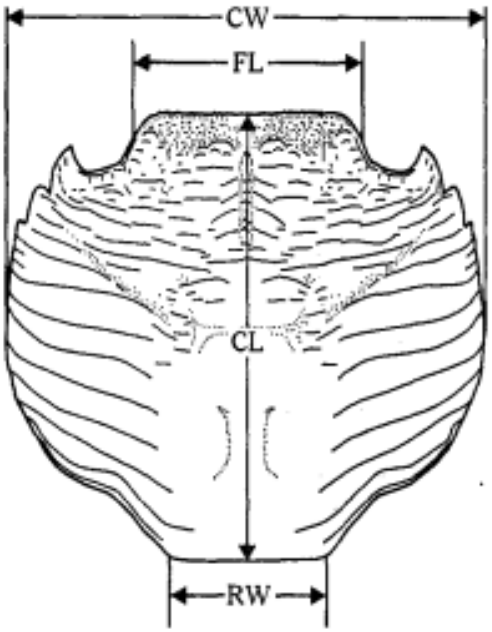

Fig. 1. Dorsal view of carapace of *Leptograpsus*, showing measurements taken. *FL*, width of frontal region just anterior to frontal tubercles. *RW*, width of posterior region. *CL*, length along midline. *CW*, maximum width. The body depth was also measured; in females but not in males the abdomen was first displaced.

Figure 1: Diagram of Crab Measurements.

The proposal should be no longer than 6 pages. The proposal will be graded on a pass / fail basis. The proposal will be evaluated on the following:

• The appropriateness of the data set. – The instructor should be able to access the data set easily. – The data set should be formatted in a manner where each row represents an observation, and each column represents a variable. – Please note that regression methods assume the observations are independent. One way of assessing whether your observations are independent is to ask if the order of the rows in your dataframe matters. If you can scramble the order of the rows without affecting any structure in your data, then your observations are likely to be independent. If scrambling the rows upends the structure of the dataframe, then your observations are not independent, and regression methods are not meant to handle dependent data. – For the categorical response variable, be sure it is binary. We have not covered enough material to tackle categorical response variables with more than 2 classes.

• The objectives are appropriate for a project that spans 4 weeks. The instructor will assess if you have at least one question involving linear regression and at least one question involving logistic regression.

Feedback will be provided for group proposals that are rejected. Groups will have the option of submitting one revised proposal (but given the relatively short time you have to complete the project, this should be avoided).

Submission Please submit your group's proposal (.pdf file) via Assignments (1 upload per group).