Final Exam

Advanced Methods for Data Analysis (36-402)

Due Friday May 10, 2019, at 11:59 PM

Remember to read the guidelines regarding the preparation and submission of a data analysis report. In particular, you are **not** allowed to collaborate with any other students regarding this work.

Do not forget the highlighting and numbering step. If you do not know what I am talking about, then read the guidelines. We expect to see a clearly written report; be sure to leave enough time for writing.

Your report should be **at most 10 pages in 12 points font and 1.5 line spacing**, including relevant plots with detailed captions; no R output. This is a page limit; do not worry if you write less. Do not report everything you do.

Any pages above the page limit will not be considered for grading.

Your data exam (report and code) must be submitted through Canvas by 11:59 PM on Friday, May 10.

Important announcements and clarifications regarding this exam will be made via Canvas. Make sure you are reading your Andrew email on a regular basis.

Experiment description

Population geneticists consider "clines" particularly favorable situations for investigating evolutionary phenomena. An example of a cline is a region where two color types (more properly denoted as morphs) of one species arrange themselves at opposite ends of an environmental gradient, with increasing mixtures occurring between. Such a cline exists near Liverpool, England, where a dark type of a local moth has flourished in response to the blackening of tree trunks by air pollution from the mills. (A moth is similar to a butterfly.) The moths are nocturnal, resting during the day on tree trunks, where their coloration acts as camouflage against predatory birds. In Liverpool, where tree trunks are blackened by smoke, a high percentage of the moths are of the dark morph. One encounters a higher percentage of the typical light (pepper-and-salt) morph as one travels from the city into the Welsh countryside, where tree trunks are lighter.

Bishop used this cline to study the intensity of natural selection (survival of the fittest). Bishop selected seven locations progressively further from Liverpool and more rural. At each location, Bishop chose eight trees at random. Equal numbers of light and dark moths were glued to the trunks in lifelike positions. (The moths involved in the experiment were dead.) One-quarter of the moths were placed on the northern, eastern, southern and western exposures of the trees. After 24- hours, a count was taken of the numbers of each morph that had been removed; the missing moths were presumably eaten by birds.

Bishop recorded orientation because he suspected that the coloration of the tree trunks varied by direction. On the northern exposure the trunk of the tree is often covered with lichen growth, which is light colored. This growth tends to mask the effects of pollution. Lichen does not usually grow on the other exposures.

Data

The file DA-exam-data.txt contains a data frame with the seven locations selected by Bishop and their distances from Liverpool. The other columns of the data set record the type of moths, the number placed and removed on the eight trees of each location, and the exposure of the side where moths were placed. Information on each column follows.

- *location*: name of the area selected;
- *dist*: distance from Liverpool in miles;
- *morph*: type (light or dark) of moth glued to trunk;
- *side*: exposure of the side of tree where moths were placed (variable added by a department faculty);
- *placed*: number of moths placed;
- *removed*: number of moths removed.

The Goals

The question of interest is whether the proportion removed differs between dark morph moths and light morphs and, more importantly, whether this difference depends on the distance from Liverpool. If the relative proportion of dark morph removals increases with increasing distance from Liverpool, that would be evidence in support of survival of the fittest, via appropriate camouflage.

Your task is to build stable, interpretable, theoretically valid models to address these research questions. Notice that this is a planned experiment and our focus is inference, not prediction.

Specifically, you must address the following issues in your report along with all other issues necessary for describing and justifying a statistically valid analysis:

Introduction

Clearly state the research questions and objectives of your study. Briefly mention your final findings.

Exploratory Data Analysis

Part 1

Describe the variables in the data set, adding only necessary univariate EDA. Identify and specify your response variable and its distribution. *Hint:* more than one column in the dataset might be needed to define your response variable.

Part 2

Do multivariate EDA (e.g. tables of counts, stacked barplots, side-by-side scatterplots…). Consider potential interaction effects. Describe any trends or interesting features that you see. Remember to analyse the relationship between the controlled factors and the response (note that your response variable is not a pre-constructed column of your dataset).

Modeling & Diagnostics

Part 3

Construct two statistical models in order to answer the research questions, one parametric and one nonparametric, *both additive*. Specify the model and its components. *Note:* pay attention to categorical and continuous variables and how to include them in the model. Explain your choices for how you code each variable; mention potential transformations and whether you decide to discretize variables. Explain how you choose your covariates, indicating which ones you are interested in, which ones are potential cofounders and potential interactions. *Hint:* remember to take into account that different locations had different total numbers of placed moths.

Part 4

To choose among the two models you proposed in 3, use a *nonparametric* test. Perform model selection and write your final model with all parameters.

Part 5

Present model diagnostics on your model. Discuss possible improvements and modifications to your model to address any violations of the model assumptions. Are there outliers?

Results

Part 6

Using the selected model, verify if the proportion of removed moths is different betwen dark and light moths. Make sure you clearly state your null and alternate hypotheses, your test statistic, how you perform your test and the assumptions you made.

Part 7

Using the selected model, verify whether a potential difference between dark and light moths depends on the distance from Liverpool. Make sure you clearly state your null and alternate hypotheses, your test statistic, how you perform your test and the assumptions you made. In the case there is a difference, does the relative proportion of dark morph removals increases or decreases with increasing distance from Liverpool?

Conclusions/Discussion:

Part 8

Discuss your results with respect to the research hypotheses. Summarize your main findings in the analysis. Discuss possible reasons for these findings. Can you make any causal statements? Make some recommendations for future work or studies but be brief.