Stat 311 Final Group Project

This group project is based on the data set Fish.csv. This data set is from a project I have been working on for the last couple of years. The data have been deidentified to the satisfaction of the client. What I can tell you is that this study takes place on a smaller river in western Washington and the fish species of interest is steelhead. After steelhead hatch further upstream from the study sites, these fish spend from two to three years (sometimes more) hanging out in the river and maturing before making their way to the ocean (some fish may never head to the ocean). Each year, in September, field biologists capture a random sample of fish at each fixed site on the river. Biologists check for a previous tag, tag untagged fish, and then weigh, and measure their lengths before releasing the fish back into the river. The biologists make an initial guess as to the age class of each captured fish (variable not included as many are missing).

While this study has multiple objectives, for this project we will focus on the following two things.

1. The relationship between weight (Weight) and length (Length).
2. Differences in key variables among length (Length), probable age class (AgeP), and site (Site).

The description of the variables for this data set are in the file FishDataDictionary.pdf.

See the last two pages of this handout for code, and report guidelines and requirements. Each group must complete Problems 1 – 4 and **one** of either Problems 5 or 6. Thus, each group will submit a report that includes five problems. There are 82 total points, with 72 points identified below for each part of each problem and up to 10 points for a readable .Rmd file, following the formatting guidelines and other requirements, including page limits, table and figure labeling, and basic grammar/readability.

There is no extra credit for doing all six problems. If you submit both problems 5 and 6, we will only read the first one and will deduct points for not following directions.

Problem 1. Reading in and preparing the data. (1+2+4 points)

a) Read in the data set. Convert all categorical variables to factors. Report the total number of observations in the data set.
b) Each year, the sampled fish are tagged before they are measured, weighed, and released. In some years, starting in year 2, a few previously tagged fish are recaptured. You can tell which fish are recaptures by looking at the variables YrCaught and YrTagged. If the year tagged is prior to year caught, then the observation is a recapture. To maintain independence of observations in our data set, we will only keep observations for the first year a fish was captured. In a sentence, explain how you can filter these data to get the desired subset. Write code to get the proper subset of data that has only the first observation per fish. This subset is the data you will use for the rest of the problems. How many observations are in the final subset of data?
c) Make a table that shows the number of observations and number missing values for each level of each of the categorical variables, and for the continuous variables (NA means missing data in R)? Omit the YrCaught and YrTagged variables.

Problem 2. Describing the data for this project. (4 points)

In your own words, using what you know about the data set, write a one paragraph description of the data that you could use for the data section of a paper. This is not an analysis of any of the variables. This is a description of what data you have available for analysis. It should include types of variables, information about sampling, and comments about observations that you choose to omit from the analysis.
[Max 100 words].

Problem 3. Simple EDA. (6+4 points)

a) Do a simple univariate exploratory data analysis for the three variables AgeP, Weight and Length. Describe the distributions for these three variables. Include appropriate summary statistics in tables and appropriate figures to support your observations. [Hint: be thoughtful on what you report as you have a page limit; e.g., you may make more graphs than what you choose to include in your report]

b) Describe the bivariate relationship between Weight and Length (use Weight on Length). Include an appropriate graph to support your findings.

Problem 4. Modeling the relationship between fish weight and fish length. (4+5+6+3+5+6 points)

a) In fisheries science, it is known that the relationship between fish weight (W) and length (L) follows the multiplicative relationship $W = aL^b$. We can linearize this relationship as $\log W = \log a + b \log L$, where log is log base 10. Since we have not used log base 10 transformations, we will use natural log transformations. The transformed relationship will be the same; only the scaling of the axes will be different. Make a scatterplot of $\ln W$ on $\ln L$. Describe the relationship you observe.

b) Fit a linear regression model for $\ln W$ on $\ln L$. Write out the regression equation and interpret the estimated slope parameter in the context of the problem.

c) Perform standard diagnostics for the regression in (b) to help determine if the assumptions for inference are met. Include a residual plot and a histogram of the residuals. Summarize your findings, identifying any concerns for assumptions that may not be met.

Regardless of your findings in part (c), assume that all assumptions for inference are met for parts (d) – (f).

d) Write out the statistical hypotheses, using symbols, to test that the slope parameter is greater than zero. Report the appropriate numbers from the R output in part (b) to evaluate this claim. For this hypothesis test, using a 5% significance level, what is the decision and what is the conclusion in the context of the problem?

e) For many fish species, the slope of the regression for a log-log model of weight on length will be close to 3. Use the information in the R output to test the claim that the slope parameter is different than 3 at the 5% significance level. Include the hypotheses using symbols, the test statistic, p-value, decision, and conclusion in the context of the problem. [Hint: this problem requires more than just reading a line of output.]

f) Use the regression model from (b) to find the expected mean weight of a juvenile steelhead that has a fork length of 70 mm. Find both a 95% confidence interval for this mean and a 95% prediction interval for a randomly selected steelhead with a fork length of 70 mm. Report the intervals in original units and interpret them in the context of the problem. [Note: the mean you are reporting in original units will be a geometric mean, not an arithmetic mean. This is a result of fitting in ln space and back transforming.]

Problem 5. Inference on the relationship between fish length and sites. (3+4+7+2+4+2 points)

a) Make comparative boxplots of fish length by site. Describe what you see.

b) Site 1 is furthest upstream while Site 3 is furthest downstream. Make two subsets of the data, one for Site 1 data and the other for Site 3 data. Using R, report and interpret the 95% confidence interval for the difference in mean lengths between the upstream and downstream sites. [Use Site 1 – Site 3]

c) Use R to test the claim that there is a difference in mean length between the most upstream and most downstream site at the 5% significance level. Include the hypotheses using symbols, the test statistic and df, p-value, decision, and conclusion in the context of the problem. [Use Site 1 – Site 3]

d) What assumptions did you make for the inference in parts (b) and (c)? Were the assumptions met or do you think any assumptions were violated? Explain. [Hint: you may want to use what you observed in the EDA in problem 3 to help answer this question]

e) What type of statistical error could you be making for the test in part (c). Explain what this error means in the context of the problem. Do you think this type of error is important in the context of the problem?

f) Do you think that the estimated population difference in the mean lengths between the upstream and downstream sites has practical significance? Try to give a reasonable answer even if you do not know much about fish biology.

Problem 6. Inference on the relationship of a Probable Age Class of 2 and Sites. (3+4+7+2+4+2 points)

a) Make a table showing the counts for Site by AgeP, where Site is in the rows. Summarize the table information in one or two sentences. Remember that AgeP is a model-based age class estimate for each fish.

b) Using R, report and interpret the 95% confidence interval for the difference in the proportion of AgeP 2-year-old fish between the upstream and downstream site. [Use Site 1 – Sites 3]

c) Use R to perform a hypothesis test for the claim that the proportion of AgeP 2-year-old fish at the most upstream site is greater than the proportion of AgeP 2-year-old fish at the most downstream site using $\alpha = 0.05$. Include the hypotheses using symbols, the z test statistic, p-value, decision, and conclusion in the context of the problem. [Use Site 1 – Sites 3]

d) What assumptions did you make for the inference in parts (b) and (c)? Were the assumptions met? Explain.

e) What type of statistical error could you be making for the test in part (c). Explain what this error means in the context of the problem. Do you think this type of error is important in the context of the problem?

f) Do you think that the estimated population difference in the proportion of AgeP 2-year-old fish between the upstream and downstream sites has practical significance? Try to give a reasonable answer even if you do not know much about fish biology.

Stat 311 Final Group Project

Project Checklist

This project is both an analysis exercise and a writing project that allows you to use and demonstrate your understanding of the ideas and methods we have covered this quarter. Use R for all the computing: plotting, key summary statistics, confidence intervals, and test statistics/*p*-values for hypothesis tests, and regression. While choosing the correct summaries and methods is important, the presentation and interpretation of the results are just as important as producing graphs and numeric output.

This report cannot exceed **five** pages for the problem answers, including any tables and figures—choose wisely to tell a story, only including what you need to support what you are talking about. You should also include a cover sheet that has your group number, the first and last names of each group member, and includes a concise description indicating how you worked together/who worked on which problems.

Please refer to the checklist below when doing this project and creating the final report.

- ☐ We did the writeup by problem number in order, with headers for each problem and parts of a problem added to the report.
- ☐ We did not cut/paste any R output into our final report. We typed in relevant output that should be included as part of the answer.
- ☐ We used reasonable rounding for numeric summaries when referring to the numbers in any written discussions/interpretations.
- ☐ We made use of `par(mfrow=c(rows, cols))` or `ggarrange` to put multiple related plots into a single figure where appropriate.
- ☐ We fully labeled all axes on graphs, including units if applicable.
- ☐ We interpreted all confidence intervals in the context of the problem.
- ☐ For all hypothesis tests, we made sure to include the null and alternative hypotheses using symbols, the test statistic, and degrees of freedom for t-tests, the *p*-value, our decision (reject the null or fail to reject the null) and an interpretation in the context of the problem, including units as appropriate.
- ☐ We used meaningful subscripts, such as μ_D for mean climate sentiment for students that identify as democrats, or we used μ_1 and μ_2 but defined what group belongs to 1 and 2.
- ☐ We made sure to use the correct=FALSE argument for any calls to `prop.test`.
- ☐ For any free response writing and for interpretations, we developed thoughtful responses by focusing on what we considered to be the key features (the TAs and I do not want to read "brain dump").
- ☐ We included the first and last names of all group participants on the cover page of the final write-up. **The TAs and I are assuming that if your name is on the paper, you read/approved of the final product.**
- ☐ We included a concise description indicating how we worked together/who worked on which problems, etc. on the cover page.
- ☐ We included the proper use of Table/Figure labels and captions and referred to included Tables/Figures in our text. We labeled all Tables and Figures consecutively in the order they appear in the report. [Note: Tables and Figures get their own consecutive labeling.]
- ☐ We did not exceed the 5-page limit (excluding the cover page) that includes all text, tables, and graphs.
- ☐ We used a font size of 11 or 12 pt.
- ☐ One group member uploaded an html version of our commented knitted R Markdown file to Canvas by the 12/10 11:30 PM PST deadline.
- □ One group member uploaded our final report as either a .doc, .docx, or .pdf file to Canvas by the 12/10 11:30 PM PST deadline.

Other Guidelines

- You are welcome to split up the work in any way that works for your group. You can divide and conquer, assign a couple of people to work on each problem, work on all problems together, or any combination you decide on. Try to make the best use of each group member's strengths. Remember, all group members get the same score.
- You can save graphs as image files to insert into MS Word or other. Simply right click on a graph in the R output and use Save image as…The default is a .png file type. You can also change to .jpg if you prefer.
- Refer to the journal article posted with this assignment for how to use Table captions (top of tables) and Figure captions (bottom of figures), and how to reference them in text. This paper tends to put the Figure or Table reference in parentheses. Other options include something like, "The summary statistics for length and weight are shown in Table 1." It is up to you and a combination of either inclusion is fine.
- Your submitted knitted html code file should contain headers for each problem and parts of problems that require code. You do not, however, include interpretations and non-code parts in your code file.
- We recommend that each group comes together for a final review of analysis methods/outputs and final edit of the writing. You should produce a project report that sounds like it was written by one person/one voice **(there should not be any sentences starting with I)**.