Stat 311 Winter 2022 Group Project

This group project is based on the data set Birds.csv. This data set is a subset of observations from a project I worked on that had the broad objective of determining which weather variables may relate to bird migration and flying heights. The study area is in New Jersey and is near a newly installed power transmission line. Although rare, bird collisions with anthropogenic structures are a concern.

The data set has observations from both fall and spring migrations. The dependent variable is average flight height in meters (Height). There are several quantitative predictors and one categorical predictor (WindDir) in addition to Season. The description of the variables for this data set are in the file BirdsDataDictionary.pdf.

See the last page of this handout for project guidelines and requirements [**Read this before starting**]. We have provided the file GroupProjectTemplate.Rmd as a template to get you started. We provide the code to read in and set up the initial data object. We also provide code for initial scatterplots and correlations for Problem 6. You will knit your final Rmd file to html, but this time you will save your html file to pdf and upload the final pdf to Gradescope [one upload per group].

Problem 1. Group Work (4 points)

Briefly describe how your group worked together. If you used divide and conquer, for each problem list the primary group member responsible for that problem.

Problem 2. Population (3 points)

The data set has a sample of observations on migratory bird flights in an area with a newly installed power transmission line. What population, if any, does this sample extend to. Explain [max 100 words].

Problem 3. Season and Wind Direction (13 points)

a) Explore the relationship between season and wind direction. Summarize the data using two contingency tables, one with raw counts (1 point) and one with percentages conditioned on season (2 points), and one graph (2 points) of your choice that will help you answer the question, do season and wind direction appear to be independent or dependent? Explain your answer based on your observations from the summary tables and graph. (3 points)
b) Estimate the difference between the proportion of crosswinds in fall and spring (2 points). Report and interpret the interval in the context of the problem (3 points).

Problem 4. Explore Average Flight Height (12 points)

Use histograms (3 points), box plots (3 points) and summary statistics (3 points) to characterize the average flight heights for all the data together, and separately by season. Compare and contrast the observed sample distributions for the Height variable for all data, and the fall and spring subsets (3 points).

Problem 5. Estimation for Average Flight Heights by Wind Direction (12 points)

In Problem 2 you looked at the relationship between season and wind direction. We will now explore flight heights within season and wind direction. Estimate the following:

a) The difference in mean average flight height for birds flying into a headwind. Report and interpret the interval in the context of the problem (3 points).
b) The difference in mean average flight height for birds flying with a tailwind. Report and interpret the interval in the context of the problem (3 points).
c) The difference in mean average flight height for birds flying with a crosswind. Report and interpret the interval in the context of the problem (3 points).
d) Summarize what these three intervals tell you about the relationship between average flight height in fall and spring, and wind direction (3 points)

Problem 6. Flight Heights in Fall and Spring (16 points)

a) Some people claim that average fall flight heights are higher than in spring for migratory birds in this region. Set up and write out the statistical hypotheses to test this claim (2 points). Provide any decisions you had to make to run the test (and why), relevant statistics and your decision in your summary (5 points). Provide interpretations of the results in the context of the problem (2 points).

b) What assumptions are you making when testing the claim in part (a) (3 points).

c) To hedge against the violation of some assumptions for the test in part (a), run a permutation test to test the same claim from part (a). Your output should only include 10 rows of the permutation output, a histogram of the null distribution with a line that locates the observed value of the difference, and the *p*-value (3 points). Do you get a result like that for part (a)? Explain (1 points).

Problem 7. Linear Regression Model for Average Flight Height (22 points)

We have provided code for scatterplots and correlations of average flight height all data on the quantitative weather variables ADP, WindSpd, and ABP. Copy/paste/edit for spring and fall, then run this code to look at the relationships, but do not include output from this code in your final submission (too many plots). Graphs are informal (no full axis labels) for you to look at relationships.

a) Considering your results for Problems 3 and 5, and the scatterplots/correlations you looked at as part of this problem, do you think fall and spring data should be used together in a single regression model or should separate models be fitted for each season? Explain [max 100 words] (3 points).

b) Using the three quantitative variables listed in the problem description, fit simple linear regression models to predict average spring flight height and pick the "best" model among the three models using all available data [no transformations and do not remove any observations]. Only report a summary of the `lm` output (1 point), a scatterplot showing the regression line for the "best" model (2 points) and diagnostics for your "best" model (4 points) [comment out or remove the code for the other models you considered]. Explain why you think the model you chose is "best." [e.g., what criteria did you use and how did it help you make your choice?] (4 points).

c) Write out the regression equation for your "best" model (1 point). Interpret the estimated slope in the context of the problem (1 point).

d) For your regression model from part (b) report and interpret in the context of the problem the confidence interval for the slope parameter (1 point).

e) For your regression model from part (b) report and interpret in the context of the problem the confidence interval for the mean average flight height when your predictor is at its Q3 value (1 point).

f) For your regression model from part (b) report and interpret in the context of the problem the prediction interval for the average flight height of a randomly selected spring migratory bird in this region when your predictor is at its Q3 value (1 point).

g) Spring flight heights have a few high outliers. Run a new linear regression model using the same predictor from your "best" model that excludes all average flight height greater than 500 m. Do you think the new model is better after removing the high outliers? Explain (3 points)

Problem 8. Followed Directions (6 points)

You do not include this problem in your submission. This problem is worth up to 6 points for following assignment directions. [Organized, clearly written and within word limits, only showed relevant output, put graphs in a single figure when appropriate, sounded like one voice, and indicated what page each problem started on according to the assignment outline when uploading to Gradescope.

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. Do all your work in an Rmd file. Unless otherwise specified, there should be some writing outside of the code chunk for each part of each problem.

You should do your best to make your output concise. This means choosing wisely when making graphs, making use of putting graphs in a single figure if they are part of the same problem, and following any given word limits.

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 Rmd file.
- □ We did not just have R output for a problem, unless indicated this was okay. We typed in relevant output after each code chunk to answer the given question(s).
- □ 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. Remember, however, that you cannot easily arrange a mix of base R and `ggplot2` graphs in one figure.
- □ We selected a reasonable number of bins for histograms done using `ggplot2`. If you used `hist` from base R, this is usually not as much of an issue.
- □ 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 <u>in words and in symbols</u>, 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 used `t.test` for all tests or confidence intervals for means.
- □ 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/essential information (the TAs and I do not want to read “brain dump”).
- □ We used the spell checker in RStudio.
- □ We included the first and last names of all group participants as authors in the Rmd file. **The TAs and I are assuming that if your name is included in the Rmd file, you read/approved of the final product.**
- □ One group member uploaded our final knitted file as a pdf file to Gradescope by 11:30 PM PST on 11 March 2022.
- □ When uploading to Gradescope, you indicated what page each problem started on according to the assignment outline.

Other Guidelines

- Use a 5% significance level for all hypothesis tests and use a 95% percent level for all confidence and prediction levels.
- 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.
- 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 has one voice **(there should not be any sentences starting with I; use we)**.