STAA 551 - Case Study

Due Friday, October 11, 11:30pm

Students will be divided into groups of 3 (subject to the class size being divisible by 3). Data will be provided to each group, along with a brief description of the data. Each group will submit one final Case Study paper containing all of the components listed below. The written portion should be 7 pages (or less), written in the same style as a journal article, but with more detail and focus on the analysis. Each student will also independently provide, by email, a peer evaluation of their group members.

Required Elements/Grading (20 points total)

Introduction (4 points)

- Background/identification of the purpose of the study.
- State response and predictor variables.
- Identify each variable as quantitative or categorical (with levels specified for categorical variable).

Summary Statistics and Graphics (4 points)

- Do this before formal model fitting.
- Include information on data cleaning/restructuring.

Analysis (5 points)

- Description/discussion of analysis with enough detail that someone else could recreate your results.
- You are encouraged to try different approaches, but please restrict yourselves to the methods we have covered in this class (no other methods, such as loess or GLMs, should be used).
- Justify any choices that you made as part of the analysis.
- Discuss model assumptions and include diagnostic checks (plots and quantitative checks).
- As you are working toward a final model, keep in mind all that we have covered (checking model assumptions, transformations, etc).

Results and Conclusions (5 points)

- Final model and fitted equation
- Tables of estimated coefficients/standard errors, predictive plots, etc.
- Other results as appropriate
- Interpretation and discussion
- Refer back to the purpose of the study (how will your fitted model be of benefit)

Overall Style (2 points)

- 7 pages or less (including graphs, but not `R` code or references)
- Use complete sentences and correct grammar.
- `R` code should **not** be included in main body of the report.
- While some tables and results can be taken directly from `R` output, the reader should **not** have to sift through superfluous output (reduce `R` tables down to only what is needed).

- Graphs should be clearly labeled.

`R` Code Appendix

- I will check congruence of `R` code vs written description.
- If `R` code is not included, there will be a 2 point deduction.

Data Description

The goal of this study is to construct a model for predicting the amount of prize money that a golfer will win, based on characteristics of her experience and performance. The data here is taken from the LPGA (Ladies Professional Golf Association). The variable `przrnd`, the amount of prize money won per round of golf played, is the relevant outcome.

The data is contained in the file 'LPGA.csv'. The total number of records in the dataset is 100 LPGA golfers.

Following is a list of the variables, along with variable descriptions:

Variable	Description	Data Type
Golfer	Name of golfer	string
rounds	Number of rounds played	integer
avedist	Average drive distance	numeric
pctfrwy	Percentage of fairways hit	percent
pctgrn	Percentage of greens in regulation	percent
aveputt	Average putts per round	numeric
avesand	Average sand shots per round	numeric
pctsandsv	Sand save percentage	percent
przrnd	Prize money per round	numeric

I had to google some of these terms. For example, *sand save percentage* is "the percent of sand shots that hit the green and were following by only one putt to hole out." You may have to further look up jargon like "hole out". If you are not so familiar with golf, as I am not, it's okay for the purposes of this case study.

Some notes:

- You may want to do a little research (not too much) on how the predictors above would be expected to affect the response.

- This data will require some cleaning and further investigation. Before you do anything, you need to look at plots and summaries for any potential problems. For example, if 99% of the data is "Category" A, and only 1% "Category B" for a predictor, perhaps including that predictor would be ill-advised.

- You may also want to change the variable type for some predictors. Objects that look like "character" objects to `R` may need to be changed to "factor". Some of the "integer" objects could also be "factors".