Problem set 5

due Friday, December 8 at 11:59pm

> **Estimated time: 6 hours**
>
> Allocate about **1 hour per problem**, though some will take longer than others. You may need more time if programming is completely new to you, or less if you have some experience already.

Instructions Upload your `.ipynb` notebook to gradescope by 11:59pm on the due date.

- Note that each problem will be graded according to this rubric. Solutions that include packages or functions not covered in this course will recieve a score no higher than 2.
- You may collaborate with any of your classmates, but you must write your own code/solutions, understand all parts of the problem, and name your collaborators.
- You should also cite any outside sources you consulted, like Stack Overflow or ChatGPT, with a comment near the relevant lines of code (see example below). Recycled code that has not been cited will be considered plagerism and receive a zero.

```
# code here was inspired by user2554330 on stack overflow:
# https://stackoverflow.com/questions/69091812/is-everything-a-vector-in-r
```

As always, create a new colab R notebook. Include the title "Problem set 5", your name, the date, and any collaborators somewhere at the top.

Dataset Information

You will work with the `verbs` dataset, part of the `languageR` package. This dataset is a simplified version of the `dative` dataset, which is described as follows:

> Data describing the realization of the dative as NP or PP in the Switchboard corpus and the Treebank Wall Street Journal collection.

Dative alternation refers to the fact that some sentences can have two different structures that convey the same basic meaning. In the `verbs` dataset, we are looking at ditransitive verbs (such as 'give' or 'offer') which can realize the recipient in two ways: as a noun phrase (NP) "give you a book", or as a prepositional phrase (PP) "give a book to you". We want to investigate what factors in the data (`AnimacyofRec`, `AnimacyofTheme`, `LengthofTheme`) impact whether the recipient is realized as a noun phrase (NP) or a prepositional phrase (PP) (`RealizationOfRec`).

Problem 1: Data Exploration

- Load the dataset and perform exploratory data analysis
- Examine the dataset's structure (`glimpse`), summary statistics, and handle missing values if any.
- Visualize the *pairwise relationship* between the response variable and each of the three explanatory variables using appropriate plots (scatter plots, histograms, etc.).

Problem 2: Model Specification

- Choose an appropriate type of model and explain why you've selected this type (e.g. regression, classification)
- Specify the model with an equation (e.g. $y = \beta_0 + \beta_1 x_1$). Note that you can create math equations in google colab text blocks by placing the equation between two dollar signs: `$y = \beta_0 + \beta_1x_1$`
- Specify the model in R code (e.g. `y ~ x`)

Problem 3: Model Fitting

- Fit the model you specified in R
- What does the model tell us? How is realization of NP/PP impacted by the animacy of the theme (`AnimacyOfTheme`), the animacy of the recipient (`AnimacyOfRec`), and the length of the theme (`LengthofTheme`)?
- What is the meaning of the intercept? Convert the estimate for the intercept to probability.

Problem 4: Model Accuracy

- Assess the accuracy of the model using cross-validation
- Take care to use `collect_metrics()` to report any accuracy metrics
- Add text explaining how accurate your model is, looking at the mean and standard deviation of accuracy.

Problem 5: Model Reliability

- Assess the reliability of your parameter estimates via bootstrapping
- Get a 95% confidence interval around the parameter estimates of your model.
- Visualize your bootstrapped distribution and shade the confidence interval.

> **i** If you get the error `contrasts can be applied only to factors with 2 or more levels`, try the bootstrap with a seed of 2 (`set.seed(2)`) and a reduced number of replicates `reps = 500`!

Problem 6: Model Predictions

Using the coefficients returned from your model from Problem 3, get the predicted probability of `PP` or `NP` (your choice) for the following:

- `AnimacyOfTheme` and `AnimacyOfRec` are both animate, and `LengthOfTheme` is 5
- `AnimacyOfTheme` is inanimate, `AnimacyOfRec` is animate, and `LengthOfTheme` is 3

> **i** A previous version of Problem 6 stated:
>
> > Create a visualization to plot your model's fit (`geom_smooth()`)
>
> You have a choice to do either version of Problem 6.