# Assignment 2 - ESM 244

### All parts due by 5:00pm PST on Friday 2/17/2023

*Binary logistic regression; parameter estimation with non-linear least squares; build Shiny App infrastructure*

## What you’ll submit for Assignment 2 by the due date

* 10 pts - Your knitted html (not Rmd) for Task 1 **(on GauchoSpace):**

a2\_task1\_lastname\_firstname.html

* 10 pts - Your knitted html (not Rmd) for Task 2 **(on GauchoSpace):**

a2\_task2\_lastname\_firstname.html

* 5 pts - Add your Github repository URL to the [Shiny App Information Page](https://docs.google.com/spreadsheets/d/1yaFF_qHK4CRt9qBDwsnkS1etnTkMcGpyEHIMyvA5D8s/edit?usp=sharing). **If your repository is private**, **add Casey (GitHub: oharac) as a collaborator** to your Shiny app repo. (5 pts)

The instructions below are what you’ll be graded on. If you want to explore further in your analysis tasks, or go further in your website or Shiny app, feel free to push yourself!

## Task 1 - Palmetto binary logistic regression (individual)

In a separate .Rmd, using the Florida palmetto data (palmetto.csv), use binary logistic regression to test feasibility of using variables plant height (height), canopy length (length), canopy width (width), and number of green leaves (green\_lvs) to classify whether a palmetto is species *Serenoa repens* or *Sabal etonia*.

**Data source:** Abrahamson, W.G. 2019. Survival, growth and biomass estimates of two dominant palmetto species of south-central Florida from 1981 - 2017, ongoing at 5-year intervals ver 1. Environmental Data Initiative. <https://doi.org/10.6073/pasta/f2f96ec76fbbd4b9db431c79a770c4d5>

Get the data: [**palmetto.csv**](https://drive.google.com/file/d/1BGqjhzZvb14GngmYry7dTbJYuc6UMJAa/view?usp=share_link)

**More information and metadata:** <https://portal.edirepository.org/nis/metadataviewer?packageid=edi.317.1>

For this task, produce a final, professionally formatted knitted HTML that contains the following:

1. An overview section describing the data, the question(s) to be addressed in your analysis, and a citation of the dataset.
2. A section containing 2 - 3 **finalized** (customized, suitable for a publication) data visualizations (with figure captions) in which you explore differences in height, canopy length, canopy width, and green leaves for the two species. If you prefer, combine the figures into a compound figure using {patchwork} or {cowplot}. Below your data visualizations, add a sentence or two with a takeaway from the plots, e.g., based on these plots, which predictor variables are more likely to help classify species correctly?
3. A section in which you perform binary logistic regression to determine the probability of a plant being either *Serenoa repens* or *Sabal etonia* based on several predictor variables. Perform the analysis twice, using cross validation to **compare two models**:
   1. Log odds of plant type using plant height, canopy length, canopy width and green leaves as predictor variable.
   2. Log odds of plant type using plant height, canopy width and green leaves (i.e., drop canopy length for this model)

Make sure you understand which species is the first ‘0’ factor level, and which is ‘1’ - you may want to convert to a factor first, then use the levels() function to check. Use repeated cross validation (ten-fold cross validation, repeated at least ten times - you can use functions from the {tidymodels} package to automate this, or manually perform the analysis using for-loops or {purrr} functions). Based on the results of the cross validation, describe which model performs better at classification; you may wish to compare AICC and BIC values as well to support your decision.

1. Train your selected model using the entire dataset, and create a finalized table (e.g., knitr::kable() and {kableExtra} functions) containing the binary logistic regression model results (at least coefficients, standard errors for the coefficients, and information for significance - consider using broom::tidy() to get you most of the way).
2. A section that evaluates how successfully this model would “classify” a plant as the correct species, using a 50% cutoff (e.g. if the probability is >=50% that it is species A, then it would be classified as species A). Use broom::augment() to find the **probabilities** (instead of **log-odds**) for each plant in the original dataset, then add a column for which species your model *would classify that plant as* (using a 50% cutoff) based on the included predictor variables. The outcome should be a **finalized table** showing, **for each species**, how many plants in the original dataset would be correctly classified and how many were incorrectly classified by the model, as well as an additional column with “% correctly classified”. Add a table caption above the table, and a 1-2 sentence conclusion paragraph after.

**To submit Task 1,** knit to HTML. Ensure that all messages, warnings are hidden but all attached packages are visible (setup chunk included). Code should be available if we click on the Code button (use [code folding](https://bookdown.org/yihui/rmarkdown-cookbook/fold-show.html) in your R Markdown YAML header). Upload your file to GauchoSpace.

## 

## Task 2: Parameter Estimation with Purrr - Lizards Length to Weight

## Source: Lightfoot, D. and W.G. Whitford. 2020. Lizard pitfall trap data from 11 NPP study locations at the Jornada Basin LTER site, 1989-2006 ver 37. Environmental Data Initiative. <https://doi.org/10.6073/pasta/4a6e258fb49c31e222ecbbcfd128967f>

Get the cleaned data: [**lizard.csv**](https://drive.google.com/file/d/1qfmMSoXgwwBNZkCPdzF0U6sPerT3qWRy/view?usp=sharing)

Raw data and metadata including species codes available at <https://doi.org/10.6073/pasta/4a6e258fb49c31e222ecbbcfd128967f>.

For task 2, you will use non linear least squares to estimate parameters of a length to weight model for lizard populations in New Mexico.

## Create a knitted.html saved as a2\_task2\_lastname\_firstname.html which provides a description of the dataset, the purpose of the analysis, a citation of the dataset, and answers the following questions.

1. Fit a snout length to weight model of the following form to all lizards in your dataframe.



* 1. Weight is given by W, snout to vent length by SVL, and a and b are the parameters that need to be fitted. Which strategy would be best to provide an initial guess? We could go with strategy one and look through the literature, but let’s practice our coding and math skills.
  2. Since we know the model is exponential in nature, we could log transform the data. If we do a standard OLS regression on the log transformed data, we can get approximations of the parameters from the regression coefficients

my\_guess\_model <- lm(log\_weight ~ log\_length, data = my\_df)

* 1. Using the coefficients function, we can then supply the NLS start list with the regression coefficients. Hint: Because you log transformed the data, you will have to mathematically transform the intercept coefficient to get the guess for parameter *a*.

1. Present your fitted model on a plot with female and male lizards separated by color. You should include the NLS model in kable output of the html.
2. Filter out the dataset for male Western Whiptail lizard (*Cnemidophorus tigrisatus*). Fit an NLS model to this subset. Compare the output from the species specific NLS model to the general NLS model for all species by graphing the model fits on the Western Whiptail male data. Report the RMSE for both models in the figure caption. Also in the figure caption briefly discuss which model should be used and why.

**Task 2 continued: Challenge yourself (optional!):**

1. Group by species and sex then nest the data. Use purrr::map() variants to parameterize length to weight models for every species separated by each sex. Hint: you may need to adjust the controls of the nls function!
2. Use the models to predict the data. Calculate the RMSE for each model created in part 3. Make a table showing the RMSE for each model.
3. Create a plot grid showing each species and sexes actual weight and the model prediction.
4. Calculate the percent difference in the *b* parameter between male and females. Plot the differences as a graph. Describe the implications of your result.

## 

## Task 3 - Build Shiny app infrastructure (Shiny App group)

For this task, you and your group should build out your Shiny app infrastructure for your term project. You should at least:

* **Build your general expected layout** (e.g. tabs, sidepanels, etc.) for the app - it is OK if this changes later on (this should be in a version-controlled RProject)
* **Start customizing the theme** - you can use an already existing theme, e.g. with {shinythemes}, or customize your own, e.g. with {bslib}...I recommend the latter for complete, fun control over the look of your app
* **Populate your tabs with placeholder widgets** (they don’t have to actually be called to do anything functional in the server at this point...but great if you’re already working on using them as inputs for reactive outputs)
* **Start populating the app** with basic information/images/maps/etc. relevant for a user (e.g. data summary, purpose of the app, etc.). See [term project guidelines](https://docs.google.com/document/d/1L2413--sYZLijV10klH_U0kpN9mkAqITDAJK4rqaYMw/edit) to refresh your memory on the overall project expectations.

**To submit this task:**

* Add your Shiny App Github repository URL (e.g., <https://github.com/oharac/visualizing_human_impacts>) to the Shiny App Information Page. **Make sure to push your latest version to Github!**
* **If your repository is private**, e.g., you have some proprietary information that should remain hidden from public view, **add Casey (GitHub: oharac) as a collaborator** to your Shiny app repo.

## 