Lab 8: Linear Regression

In this lab, we will learn regression approaches and ways of comparing different models. The learning goals of this lab are to:

- Run a linear regression using the `lm()` function
- Evaluation the results of a linear regression using the `plot()` function
- Consider multi-collinearity of predictors using the `cor.test()` function
- Compare different model structures using the `AIC()` function

In the first part of today's lab, your instructor will walk through the code and relevant functions with you.

In the second part of today's lab, you will be asked to complete your assignment in a new RMarkdown file. Please be sure to answer assignment questions using full sentences, including information as you would in a final report. *Submit your RMarkdown file rendered as a pdf to the class Canvas site under the Assignments folder. Lab assignments are due at 11:59pm the day before your next lab session meets.*

Setup

Create a new project for this week's lab named `week8` in the `ENV710` folder. Open and save a new R Markdown script named `week8script.Rmd` in this project. Edit the title of the script accordingly, and add a first **Setup** header to your script. Then, create and run a new code chunk to load in the packages and data you will use today.

The data for this week, "hbr_maples.csv", was collected by the Hubbard Brook Long Term Ecological Research program based at Hubbard Brook Experimental Forest in the White Mountains in New Hampshire. Here, we are interested in looking at the growth of sugar maple seedlings in two watersheds, one that is a reference watershed with no treatment and one that is watershed 1 (W1) which received calcium additions to better support sugar maple growth and offset the effects of soil acidification due to acid rain.

```
# Load necessary packages.
library(here)
library(tidyverse)

# Load necessary data.
maples <- read_csv("hbr_maples.csv")
```

Linear Regression

In the first part of today's lab, we will perform a single variable regression and become familiar with the outputs in R. Create a new header labeled **Linear Regression** in your script.

Step 1 - Define your research question.

Today, we are first interested in addressing the question, **is there a relationship between seedling height and mass?**

Step 2 - Examine your data and possible correlations.

```
# Create and examine histograms of raw values.
(hist1 <- ggplot(maples, aes(x = stem_length)) +
  geom_histogram() +
  theme_bw())

(hist2 <- ggplot(maples, aes(x = stem_dry_mass)) +
  geom_histogram() +
  theme_bw())

# Create and examine an initial scatterplot of data.
(scatter1 <- ggplot(maples, aes(x = stem_length,
                                y = stem_dry_mass)) +
  geom_point() +
  theme_bw())
```

Practice Problem: What do you notice about the distribution of raw data values? Based on your initial scatterplot, what relationship do you anticipate will appear in the linear regression results?

Step 3 - Fit regression model.

```
# Fit linear regression model with one dependent variable (mass) and one
# independent variable (length).
maple.lm1 <- lm(stem_dry_mass ~ stem_length, data = maples)

# Examine the results.
summary(maple.lm1)
```

Practice Problem: Using the appropriate R-squared value given your model structure, what percentage of variance in your data is explained by this model?

Step 4 - Evaluate model diagnostics.

Remember, some of the key assumptions that must be met when performing a linear regression are focused on the model *residuals*, meaning you must first fit the model to your data and then examine the model fit.

```
# Examine model residual plots (4 in total).
plot(maple.lm1)
```

Note, since we are working in an RMarkdown, these plots will populate automatically below a given code chunk. However, if you would like to run this same code in a plain R script, you will need some additional preparatory lines:

```
windows() # quartz() if on a Mac

par(mfrow = c(2,2))

plot(linearmodel)
```

There don't appear to be any visible outliers, particularly ones that surpass the Cook's distance. However, in the very first plot, there does appear to be evidence of heteroscedasticity, with variance increasing as you progress along the x axis. So, let's employ a data transformation to see if we can help with this issue.

```r
# Log-transform mass values since these appeared skewed at the start.
maples$log_stem_dry_mass <- log10(maples$stem_dry_mass)

# Re-fit model using these values.
maple.lm2 <- lm(log_stem_dry_mass ~ stem_length, data = maples)

# Examine the results and the residuals.
summary(maple.lm2)

plot(maple.lm2)
```

Practice Problem: What changes in the model residuals plots do you notice in this newer version of the model?

Step 5 - Communicate the results.

We feel confident that the model meets the assumptions, so we communicate our findings:

The results of a linear regression suggest that sugar maple seedling height significantly predicts stem mass ($F(1,357) = 130.4$, $p < 0.001$, Adjusted $R2 = 0.27$), with seedling height displaying a significant positive relationship with stem mass ($B = 0.009$, $p < 0.001$). Prior to fitting the model, stem mass data were log-transformed. The final model fit was visually inspected to assess model residuals for outliers and homoscedasticity.

Bonus - Predicted values.

If you would like to display actual data values that exist in your original dataset as well as predicted values given your chosen model structure, you can use the `fitted()` function. For example, `fitted(maple.lm2)` will create a vector of values that are the predicted seedling masses if the model inputs are the same as the predictor variable values (and remember we *log-transformed* these at the start, so we'll need to take that into account when plotting).

```r
# Create a new column of predicted seedling mass values based on your
# model results and using the original seedling height input values.
mass_predict <- fitted(maple.lm2)

# Add these values to your original dataset for plotting purposes.
maples <- maples %>%
  mutate(log_mass_predict = mass_predict) %>% # log-transformed values
  mutate(stem_dry_mass_predict = 10^(log_mass_predict)) # un-transformed values

# Plot these results.
(scatter2 <- ggplot(maples, aes(x = stem_length))+
  # first the predicted values
  geom_point(aes(y = stem_dry_mass_predict), color = "darkolivegreen3") +
  # then the raw data
  geom_point(aes(y = stem_dry_mass), color = "black") +
  # don't forget axis labels with units
  labs(x = "Seedling Height (mm)",
       y = "Seedling Mass (g)") +
  # and this is an optional way to transform you axis
  scale_y_log10() +
  theme_bw())
```

Multiple Linear Regression

In the second part of today's lab, you will build multiple linear regression models and become familiar with how to compare their model fits. Create a new header labeled ** Multiple Linear Regression** in your script.

Step 1 - Define your research question.

In this second part of the lab, we are interested in addressing the question, **is seedling mass significantly predicted by seedling height, leaf mass, leaf area, calcium treatment, and elevation?**

Practice Problem: Before building your model, consider how the above factors might influence your results. Do you think leaf mass or area will have a relationship (positive or negative) with seedling mass? Will calcium treatment or elevation influence seedling mass? Why or why not? Add your hypotheses into your script.

Step 2 - Examine your data and possible correlations.

Using the workflow above, create figures to explore the data distribution and relationships with the dependent variable (`stem_dry_mass`) for all of the possible covariates or independent variables:

- `stem_length` (seedling height in millimeters)
- `leaf_dry_mass` (leaf dry mass in grams from 2 combined leaf samples)
- `corrected_leaf_area` (leaf area in square centimeters)
- `watershed` (W1 is the calcium-treated watershed)
- `elevation` (including Low and Mid-elevation sites)

What do you notice about the data distributions? What relationships do you anticipate with seedling mass (`stem_dry_mass`) in the linear regression model?

To investigate the possibility of multi-collinearity, use the `cor.test()` function to test for possible correlations among the continuous variables included in your model. If you are unsure how to input the data, learn more about the function by typing in `?cor.test()` into your console.

Are there any variables from your initial list that are strongly correlated with one another?

Step 3 - Fit regression model(s).

Based on our findings above, we will fit two models so that we can include a measure of leaf data, but not include variables that we found to be correlated with one another.

Using the same `lm()` function you used above, fit two multiple linear regression models named `maples.mlm1` and `maples.mlm2` with the following structures:

(1) seedling mass = seedling height + leaf mass + watershed area + elevation

(2) seedling mass = seedling height + leaf area + watershed area + elevation

Note, given your previous model results, you may choose to continue to use the log-transformed data for seedling mass to encourage your model residuals to display homoscedasticity.

Use the `summary()` function you use above to examine the results of both models. Do the coefficients for the independent variables in each model make sense given your initial hypotheses and data exploration?

Step 4 - Evaluate model diagnostics.

Examine the model residuals for both models using the `plot()` function you learned above. Notice how there is a significant outlier in the first model results.

To remove this outlier, you may use the following structure to create a new dataset without this row of data. Here, the example is removing the 4th row.

```
new_data <- old_data[-4,]
```

Remove the necessary outlier from your dataset and re-fit your two linear regression models. Then, re-evaluate your model fits using the `plot()` function. Do you now feel confident that your model residuals meet the necessary assumptions?

Since you have more than one model, to compare how well they each describe the variability in the data, you may use Akaike Information Criterion, or AIC, values to describe them relative to one another. *Remember*, AIC values are only comparable between similar model structures and models that have been fit *to the same data*. **Lower values indicate a better model fit.** You may use the following code to save the AIC values for the models you have created.

```
lm1_AIC <- AIC(lm1)
```

Which model do the AIC values suggest displays a better fit?

Note, here we are using AIC values to compare between models that use the same underlying data and differ only in the variables included. Another reason people frequently use AIC values is to trim down the number of variables in a model. If this is your goal, you must think critically about the step-wise removal of any variables.

Step 5 - Communicate the results.

Based on the example text from the previous linear regression model you ran and the sample text provided in class, report the results of your selected multiple linear regression model in text form, including R2 values and measures of significance for the overall model in addition to covariate coefficient values and measures of their individual significance. Take care in how you report continuous versus categorical variables.

Bonus - Interaction terms.

In some cases, you may have variables that do not display multi-collinearity, so they can be used in the same model, but you may have some underlying information about the system that suggests that the two variables may have a relationship for other reasons. For example, you may hypothesize that the calcium treatment in Watershed 1 might have a positive effect on seedling height (i.e., the addition of calcium decreased soil acidity enough to support faster plant growth), even if your data doesn't display a relationship. When this is the case, you may include two variables as an *interaction term* in the model. The model will then estimate a coefficient for the first term, the second term, and their interaction. To see what this looks like, fit the following model:

seedling mass = (seedling height x watershed area) + leaf mass + elevation

Using the `summary()` function, examine the coefficients that are provided. Does the interaction appear significant?

Assignment

Complete the following questions by using the appropriate analyses in a new R Markdown script. Your script should include an appropriate title, load in the proper packages and data, and be structured using helpful headers. Please take care to follow the guidelines listed above when reporting the results of your statistical tests. You will submit a pdf file that displays *all* code and outputs. Be sure to submit that script, containing your rendered figures and with any requested answers as *full sentences*, via Canvas.

For the following assignment, you will use two, publicly available datasets that have been posted to Canvas:

- `NC_Census.csv` - Census data from 20 counties in North Carolina compiled by the U.S. Census Bureau and accessible at https://www.census.gov/quickfacts/fact/table/durhamcountynorthcarolina/PST045224.
- `NC_Recreation_Acreage.csv` - Local, state, and federal outdoor recreation acreage data for all counties in North Carolina compiled by the North Carolina Office of State Budget and Management and accessible at https://linc.osbm.nc.gov/explore/dataset/outdoor-recreation-acreage/export/?disjunctive.area_name&disjunctive.variable&disjunctive.year&refine.variable=State+Outdoor+Recreation+Acreage&refine.variable=Local+Outdoor+Recreation+Acreage&refine.variable=Federal+Outdoor+Recreation+Acreage.

Your research question for this assignment is - What demographic factors significantly predict outdoor recreation acreage at the county level across 20 North Carolina counties?

(1) Load, tidy, and combine the datasets in RStudio. Filter the Office of State Budget data to include only data for the 20 relevant counties (see census data) and sum together local, state, and federal recreation acreage to form your outcome or dependent variable value of total recreation acreage by county. Filter the census data down to include only the following ten items:

- **Population estimates, July 1, 2023, (V2023)**
- **Persons under 18 years, percent**
- **Person 65 years and over, percent**
- **Female persons, percent**
- **White alone, percent**
- **Black or African American alone, percent (a)**
- **Hispanic or Latino, percent (b)**
- **Median value of owner-occupied housing units, 2019-2023**
- **Persons per household, 2019-2023**
- **Median households income (in 2023 dollars), 2019-2023**

(2) Using the workflow presented in lab, perform a multiple linear regression to investigate potential predictors of total outdoor recreation acreage. You must decide which of the initial 10 variables make sense to include based on your conceptual understanding of relevant factors, model outputs, diagnostics, and comparative model fits (AIC values). There is no "right answer" for this analysis - make an informed decision based on the data and understanding that you have.

(3) In 6-8 sentences, narrate your decision-making. Justify your inclusion of certain variables, investigation of any collinearity of variables, and model diagnostics. Did you choose to log-transform any data? Did you have concerns after exploring any of the model diagnostics? If so, how did that change your model selection? Please be sure to include information that helped you determined whether multi-collinearity, heteroscedasticity, or poor model fit was a concern.

(4) In 3-4 sentences, describe the final model output as you would for a final report.

Works Cited:

Juice, S. and T. Fahey (2019). Health and mycorrhizal colonization response of sugar maple (Acer saccharum) seedlings to calcium addition in Watershed 1 at the Hubbard Brook Experimental Forest ver 3. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/0ade53ede9a916a36962799b2407097e