Lab 6 - Modeling I

Introduction

In this lab you'll start your practice of statistical modeling. You'll fit models, interpret model output, and make decisions about your data and research question based on the model results.

Guidelines

Your plots should include an informative title, axes should be labeled, and careful consideration should be given to aesthetic choices.

In addition, the code should all the code should be be able to be read (not run off the page) when you render to PDF. Make sure that is the case, and add line breaks where the code is running off the page.

> **!** Important
>
> Render your document. If your code is running off the page for a given question such that we can't see your entire code, we will not evaluate any of the code for that question. The question will automatically receive a 0. This is something you can and should verify before you turn in your work.

You should have at least 3 commits with meaningful commit messages by the end of the assignment.

Additionally, if you're using functions that are not introduced in the course materials, you must cite your sources.

> **!** Important
>
> Failure to cite outside resources used, including Large Language Models like Chat GPT, is a violation of the Duke Community Standard and will be treated as such.

Part 1 - Smoking during pregnancy

Question 1

We are interested in the impact of smoking during pregnancy. Since it is not possible to run a randomized controlled experiment to investigate this impact, we will instead use a data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. This is a random sample of 1,000 cases from a data set released in 2014 by the state of North Carolina. The data set is called `births14`, and it is included in the **openintro** package you loaded at the beginning of the assignment.

a. Create a version of the `births14` data set dropping observations where there are `NA`s for `habit`. You can call this version `births14_habitgiven`.

b. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions. Create an appropriate plot displaying the relationship between `weight` and `habit`. In 2-3 sentences, discuss the relationships observed.

c. Now, fit a linear model that predicts `weight` from `habit`. Provide the tidy summary output below.

d. Write the estimated least squares regression line below using proper notation.

> **Tip**
>
> If you need to type an equation using proper notation, type your answers in-between `$$` and `$$`. You may use `\hat{example}` to put a hat on a character.

Question 2

a. Another researcher is interested in assessing the relationship between babies' weights and mothers' ages. Fit another linear model to investigate this relationship. Provide the summary output below.

b. In 2-3 sentences, explain how the regression line to model these data is fit, i.e., based on what criteria R determines the regression line.

c. Interpret the intercept in the context of the data and the research question. Is the intercept meaningful in this context? Why or why not?

d. Interpret the slope in the context of the data and the research question.

Part 2 - Parasites

Parasites can cause infectious disease – but not all animals are affected by the same parasites. Some parasites are present in a multitude of species and others are confined to a single host. It is hypothesized that closely related hosts are more likely to share the same parasites. More specifically, it is thought that closely related hosts will live in similar environments and have similar genetic makeup that coincides with optimal conditions for the same parasite to flourish.

In this part of the lab, we will see how much evolutionary history predicts parasite similarity.

The dataset comes from an Ecology Letters paper by Cooper at al. (2012) entitled "Phylogenetic host specificity and understanding parasite sharing in primates" located here. The goal of the paper was to identify the ability of evolutionary history and ecological traits to characterize parasite host specificity.

Each row of the data contains two species, `species1` and `species2`.

Subsequent columns describe metrics that compare the species:

- `divergence_time`: how many (millions) of years ago the two species diverged. i.e. how many million years ago they were the same species.
- `distance`: geodesic distance between species geographic range centroids (in kilometers)
- `BMdiff`: difference in body mass between the two species (in grams)
- `precdiff`: difference in mean annual precipitation across the two species geographic ranges (mm)
- `parsim`: a measure of parasite similarity (proportion of parasites shared between species, ranges from 0 to 1.)

The data are available in `parasites.csv` in your `data` folder.

Question 3

Let's start by reading in the `parasites` data and examining the relationship between `divergence_time` and `parsim`.

a. Load the data and save the data frame as `parasites`.

b. Based on the goals of the analysis, what is the response variable?

c. Visualize the relationship between the two variables.

d. Use the visualization to describe the relationship between the two variables.

Question 4

Next, model this relationship.

a. Fit the model and write the estimated regression equation.

b. Interpret the slope and the intercept in the context of the data.

c. Recreate the visualization from Question 3, this time adding a regression line to the visualization.

d. What do you notice about the prediction (regression) line that may be strange, particularly for very large divergence times?

Question 5

Since `parsim` takes values between 0 and 1, we want to transform this variable so that it can range between $(-\infty,+\infty)$. This will be better suited for fitting a regression model (and interpreting predicted values!)

a. Using mutate, create a new variable `transformed_parsim` that is calculated as `log(parsim/(1-parsim))`. Add this variable to your data frame.

> **Note**
>
> `log()` in R represents the **nautral log**.

b. Then, visualize the relationship between divergence_time and `transformed_parsim`. Add a regression line to your visualization.

c. Write a 1-2 sentence description of what you observe in the visualization.

Question 6

Which variable is the strongest individual predictor of parasite similarity between species?

To answer this question, begin by fitting a linear regression model to each pair of variables. Do not report the model outputs in a tidy format but save each one as `dt_model`, `dist_model`, `BM_model`, and `prec_model`, respectively.

- `divergence_time` and `transformed_parsim`
- `distance` and `transformed_parsim`
- `BMdiff` and `transformed_parsim`

- `precdiff` and `transformed_parsim`

a. Report the slopes for each of these models. Use proper notation.

b. To answer the question of interest, would it be useful to compare the slopes in each model to choose the variable that is the strongest predictor of parasite similarity? Why or why not?

Question 7

Now, what if we calculated R^2 to help answer our question? To compare the explanatory power of each individual predictor, we will look at R^2 between the models. R^2 is a measure of how much of the variability in the response variable is explained by the model.

As you may have guessed from the name R^2 can be calculated by squaring the correlation when we have a simple linear regression model. The correlation r takes values -1 to 1, therefore, R^2 takes values 0 to 1. Intuitively, if r=1 or -1, then R^2=1, indicating the model is a perfect fit for the data. If r 0 then R^2 0, indicating the model is a very bad fit for the data.

You can calculate R^2 using the glance function. For example, you can calculate R^2 for `dt_model` using the code `glance(dt_model)$r.squared`.

a. Calculate and report R^2 for each model fit in the previous exercise.

b. To answer our question of interest, would it be useful to compare the R^2 in each model to choose the variable that is the strongest predictor of parasite similarity? Why or why not?

Part 3 - GDP and life expectancy

Gapminder is a "fact tank" that uses publicly available world data to produce data visualizations and teaching resources on global development. We will use an excerpt of their data to explore relationships among world health metrics across countries and regions between the years 1952 and 2007. The data set is called `gapminder`, from the **gapminder** package. A table of variables can be found below.

- `country`: The country name
- `continent`: The continent name
- `year`: ranges from 1952 to 2007 in increments of 5 years
- `lifeExp`: life expectancy at birth, in years
- `pop`: population of country

- `gdpPercap`: GDP per capita (US$, inflation-adjusted)

Question 8

a. **Data:** For our analysis, we will only be working with data from 2007. Below, filter the data set so only values from the year 2007 are shown. Save this data set as `gapminder_07` and use it for the remainder of this exercise and the following.

b. **Visualization:** We are interested in learning more about GDP, and we'll start with exploring the relationship between life expectancy and GDP. Create two visualizations:

   - Scatter plot of `gdpPercap` vs. `lifeExp`.
   - Scatter plot of `gdpPercap_log` vs. `lifeExp`, where `gdpPercap_log` is a new variable you add to the data set by taking the natural log of `gdpPercap`.

   First describe the relationship between each pair of the variables. Then, comment on which relationship would be better modeled using a linear model, and explain your reasoning.

c. **Model fitting:**

   - Fit a linear model predicting log gross domestic product from life expectancy. Display the tidy summary.

d. **Model evaluation:**

   - Calculate the R-squared of the model using two methods and confirm that the values match: first method is using `glance()` and the other method is based on the value of the correlation coefficient between the two variables.
   - Interpret R-squared in the context of the data and the research question.

Question 9

Next, we want to examine if the relationship between GDP and life expectancy that we observed in the previous exercise holds across all continents in our data. We'll continue to work with logged GDP (`gdpPercap_log`) and data from 2007.

a. **Justification:** Create a scatter plot of `gdpPercap_log` vs. `lifeExp`, where the points are colored by `continent`. Do you think the trend between `gdpPercap_log` and `lifeExp` is different for different continents? Justify your answer with specific features of the plot.

b. **Model fitting and interpretation:**

- Regardless of your answer in part (a), fit an additive model (main effects) that predicts `gdpPercap_log` from life expectancy and continent (with Africas as the baseline level). Display a tidy summary of the model output.
- Interpret the *intercept* of the model, making sure that your interpretation is in the units of the original data (not on log scale).
- Interpret the *slope* of the model, making sure that your interpretation is in the units of the original data (not on log scale).

c. **Prediction:** Predict the GDP of a country in Asia where the average life expectancy is 70 years old.

Question 10

Communication is a critical yet often overlooked part of data science. When we engage with our audience and capture their interest, we can ultimately better communicate what we are trying to share.

Please watch the following video: Hans Rosling: 200 years in 4 minutes.

Then, write a paragraph (4-5 sentences) addressing the following:

- What did you enjoy about the presentation of data? What did you find interesting
- Were there any aspects of the presentation that were hard to follow? If so, what?
- What are your general take-aways from this presentation?
- What are your general take-aways from how this presentation was given?

Wrap-up

Submission

Once you are finished with the lab, you will submit your final PDF document to Gradescope.

> **Warning**
>
> Before you wrap up the assignment, make sure all of your documents are updated on your GitHub repo. We will be checking these to make sure you have been practicing how to commit and push changes.
> You must turn in a PDF file to the Gradescope page by the submission deadline to be considered "on time".

To submit your assignment:

- Go to http://www.gradescope.com and click *Log in* in the top right corner.
- Click *School Credentials* → *Duke NetID* and log in using your NetID credentials.
- Click on your *STA 199* course.
- Click on the assignment, and you'll be prompted to submit it.
- Mark all the pages associated with question. All the pages of your lab should be associated with at least one question (i.e., should be "checked").

! Checklist

Make sure you have:

- attempted all questions
- rendered your Quarto document
- committed and pushed everything to your GitHub repository such that the Git pane in RStudio is empty
- uploaded your PDF to Gradescope
- selected pages associated with each question on Gradescope

Grading

The lab is graded out of a total of 50 points.

You can earn up to 5 points on each question:

- 5: Response shows excellent understanding and addresses all or almost all of the rubric items.
- 4: Response shows good understanding and addresses most of the rubric items.
- 3: Response shows understanding and addresses a majority of the rubric items.
- 2: Response shows effort and misses many of the rubric items.
- 1: Response does not show sufficient effort or understanding and/or is largely incomplete.
- 0: No attempt.