HW 02: Ice duration and air temperature in Madison, WI

Bootstrap confidence interval for the slope

Name

2024-09-06

! Important

Due:

- Friday, September 13th

Introduction

In this weeks homework, you'll use simple linear regression to analyze the relationship between air temperature and ice duration for two lakes in Wisconsin.

Learning goals

By the end of the hw you will…

- Be able to fit a simple linear regression model using R.
- Be able to interpret the slope and intercept for the model.
- Be able to use simulation-based inference to draw conclusions about the slope.
- Continue developing a workflow for reproducible data analysis.

Getting started

- Go to RStudio and login with your College of Idaho Email and Password.
- Make a subfolder in your hw directory to store this homework.

- Log into Canvas, navigate to Homework 2 and upload the `hw-02.qmd` file into the folder your just made.

Packages

The following packages are used in the hw:

```r
library(tidyverse)
library(broom)
library(infer)
library(knitr)
library(lterdatasampler)
```

Data: Ice cover and air temperature

The datasets `ntl_icecover` and `ntl_airtemp` lterdatasampler R package[1] contain information about ice cover and air temperature, respectively, at Lake Monona and Lake Mendota for days in 1885 through 2019. The data were originally collected at a US Long Term Ecological Research program (LTER) Network.

The primary variables for this analysis are

- `year`: a number denoting the year of observation
- `lakeid`: a factor denoting the lake name
- `ice_duration`: a number denoting the number of days between the freeze and breakup dates of each lake
- `ave_air_temp_adjusted`: a number denoting the air temperature in degrees Celsius, collected in Madison, WI and adjusted for biases

Exercises

Goal: The goal of this analysis is to use linear regression understand the association between average air temperature and ice duration for two lakes in Madison, Wisconsin that freeze for a portion of the year. Because ice cover is impacted by various environmental factors, researchers are interested in examining the association between these two factors to better understand how changing air temperature is impacting the ice duration.

[1] https://lter.github.io/lterdatasampler/index.html

Write all code and narrative in your Quarto file. Write all narrative in complete sentences. Throughout the assignment, you should periodically **render** your Quarto document to produce the updated PDF.

> Tip
>
> Make sure we can read all of your code in your PDF document. This means you will need to break up long lines of code. One way to help avoid long lines of code is is start a new line after every pipe (`|>`).

Exercise 1

Fill in the code below to create a new data frame, `icecover_avg`, of the average ice duration by lake then year. How many observations (rows) are in `icecover_avg`? How many variables (columns)? As you work through this document make sure you remove all `#| eval: FALSE` statements.

```
icecover_avg <- ntl_icecover |>
  group_by(____, _____) |>
  summarize(avg_ice_duration = mean(_____), .groups = "drop")
```

Answer here.

Exercise 2

- Use `ntl_airtemp` to create a new data frame `airtemp_avg` of the average air temperature by year. Call this new variable `avg_air_temperature` Hint: the code in the previous exercise will be useful.
- Then, the pre-written code will join `icecover_avg` and `airtemp_avg` to make a new data frame `ice_air_joined` and removes the years that have missing data for the average annual air temperature or the average annual ice duration. Look at the documentation and explain how the function `drop_na` works if you don't specify the columns where you want to drop missing values (e.g. `avg_air_temperature` and `avg_ice_duration` in this case)?

```r
# Make airtemp_avg here

ice_air_joined <- icecover_avg |>
  left_join(airtemp_avg, by = "year") |>
  drop_na(avg_air_temperature, avg_ice_duration)
```

Answer here.

> Important
>
> You will use `ice_air_joined` for exercises 3 - 7.

Exercise 3

Make a histogram to visualize the distribution of `avg_ice_duration` and calculate summary statistics for the center and spread of the distribution. Include informative axis labels and an informative title on the visualization.

Add code block and histogram here.

Use the visualization and summary statistics to describe the distribution of `avg_ice_duration` . Include the shape, center, spread, and potential presence of outliers in the description.

Insert answer here.

Exercise 4

Make a histogram of the distribution of `avg_air_temp`. Then make another visualization of the distribution of `avg_air_temp` using a different type of plot. Include informative axis labels and an informative title on both visualizations.

Add code block here.

What type of plot did you make? What is one feature of the distribution that is more apparent in the histogram than in the other plot? What is one feature of the distribution that is more apparent in the other plot than in the histogram?

Write answer here.

> Tip
>
> This is a good place to render changes to your PDF.

Exercise 5

Make a visualization of `avg_ice_duration` versus `year` with the points colored by `lakeid`. Write two observations from the visualization.

Add code block here.

Exercise 6

Create a visualization of `avg_air_temp` versus `year`. Include informative axis labels and an informative title on the visualization.

- Use the visualization to write two observations about the trend of average air temperature over time.
- Based on the visualizations of average ice duration and average air temperature over time, would you expect the linear model describing the association between average ice duration and average air temperature to have a positive or negative slope? Briefly explain.

Add code block here.

Write written answers here.

Exercise 7

Researchers would like to use a linear model to understand variability in the average ice duration based on the average air temperature. **Their analysis will focus only on Lake Monona.** Write the form of the statistical model the researchers would like to estimate using the template below. Use mathematical notation (i.e. with greek letters) and variable names (`avg_air_temp` and `avg_ice_duration`) in the equation.

$$Response = intercept + slope \times predictor$$

Tip

Click here for a guide on writing mathematical symbols using LaTex. You will need to use a backslash (\) before each underscore in the LaTex code. For example, `avg_air_temp` will be written as $avg\backslash\_air\backslash\_temp$.

Tip

This is a good place to renderyour quarto document.

Exercise 8

Fit and display the output of the regression model corresponding to the statistical model in the previous exercise. **As in the previous exercise, only observations from Lake Monona should be included in the analysis.** Use the `tidy` and `kable` functions to neatly display the model output using three decimal places.

- Write the equation of the fitted model. Use mathematical notation and variable names (`avg_air_temp` and `avg_ice_duration`) in the equation.
- Interpret the slope in the context of the data.
- Does the intercept have a meaningful interpretation in this context? If so, interpret the intercept in the context of the data. Otherwise, explain why not.

Add code block here.

Add written answers here.

Exercise 9

Use bootstrapping to construct a 93% confidence interval for the slope for Lake Monona. Follow these steps to accomplish this:

- First, set a seed for simulating reproducibly. Use the seed `212`.
- Save the value of slope estimated from the data.
- Then, simulate the bootstrap distribution of the slope using 1,000 bootstrap samples.
- Visualize the bootstrap distribution.
- Calculate the bounds of the confidence interval using the percentile method.
- Interpret the confidence interval in the context of the data.

Add code block here.

Add written answer here.

> Tip
>
> Bootstrapping can take some time. You'll notice it make take more time to render your document once you've added your bootstrapped confidence interval.

Exercise 10

> There is a statistically significant linear relationship between average air temperature and average ice duration on Lake Monona ($\beta_1 \neq 0$).

Does the confidence interval you calculated in the previous exercise support or refute this claim? Briefly explain.

Written answer here.

> Tip
>
> Render your document here.

Submission

We'll be submitting PDF documents and .qmd files to Canvas.

> ⚠ Warning
>
> Before you wrap up the assignment, make sure you have rendered your document and that the PDF appears as you want it to.

To submit your assignment, upload the .qmd and PDF files to Canvas. Please unzip them before uploading them.

Grading (50 pts)

Component	Points
Ex 1	2
Ex 2	4
Ex 3	5
Ex 4	5
Ex 5	4
Ex 6	5
Ex 7	3
Ex 8	6
Ex 9	8
Ex 10	4
Workflow & formatting	4[2]

Component	Points

[2] The “Workflow & formatting” grade is to assess the reproducible workflow and document format. This includes having a neatly organized document with readable code and your name and the date in the YAML.