Lab 9 – Modeling Melanoma Rates

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please submit your lab report as a **pdf** to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.
- Be sure that all **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview

- Melanoma rates in different U.S. states vary—but what might explain that variation?
- The *Melanoma.xlsx* file contains melanoma rates, in addition to other potentially related variables from various sources (most of which is from 2014-2017). Note this data includes the 48 contiguous states (excludes Alaska and Hawaii).
- We'll also be investigating a claim made in the cancercenter.com blog linked later regarding possible risk factors for melanomas. Does their claim stack up when we look at the data?

STEP 0

- **Pre-lab work**
  - Complete the pre-lab tutorials for Lab 9 first: https://stat212-learnr.stat.illinois.edu/
- **Download** the Melanoma.xlsx file to your computer and then **import** into your RStudio session.
- Create a new **R script** (or use the **RMarkdown file** if you are using that option)
- We'll be using tidyverse functions again for this one…do you remember what you should run first??

Variables

Name: State Name

abb: State Abbreviation

mel: Number of new melanoma cases per 10,000 people each year

sun: The average sun exposure in each state, measured in kilojoules per square meter (kJ/m^2)

ocean: Whether the state borders an ocean or not

temp: Average daily high temperature year around

pop: State Population

density: Average number of state residents per square miles

age: Average age of state residents

white: Proportion of state residents that identify as both "white" and "non-hispanic"

income: Average income of state residents

https://www.cancercenter.com/community/blog/2016/05/unhealthiest-states-for-skin-cancer-risk-may-surprise-you

This article linked above is considering different possible factors that might explain why some states have higher melanoma rates. One particular factor suggested by Dr. Charles Komen Brown is that **states with lower sun exposure on average tend to have higher melanoma rates.** He offers one possible *causal* explanation by saying *it is because residents in this state are not used to thinking about sun exposure and skin protection*.

(5pts) Question 1: Create a scatterplot to check whether there is a correlation between these variables. Assign melanoma rate as the response variable (y axis). Also, in place of using geom_point, *substitute* in the following geom code to instead plot each state's abbreviation in the coordinate space:

```
geom_text(aes(label=abb),hjust=0, vjust=0, size = 3, fontface = 2)
```

In addition:

- Add a best fit line (Standard error shading optional, Color choice up to you)
- Add an appropriate title
- Adjust the x and y axis titles to be fully written in 1-3 words or units (rather than the default variable abbreviations)
- Use the theme_classic() theme style (to provide a blank background)

Please include an image of this scatterplot in the report (code optional)

Which states have higher melanoma rates—those with more sun exposure on average, or less sun exposure?

Which one or two states appear to have the largest residuals in this model?

(5pts) Question 2: Now, create a simple linear model to see how much evidence we have that state average sun exposure levels correlate with melanoma rates, and how much variability it explains.

Copy or screenshot the <u>summary output</u> of this model into your report *(starting with the row that says "Coefficient" through to the end).*

Interpret the multiple r-squared value *(Perhaps try the fill in the blank template we learned in chapter 13).*

Is there evidence that average state sun exposure and state melanoma rate have at least some linear association? *(Don't just say "reject" or "fail to reject" the null hypothesis—use the interpretational language we learned in Chapter 2: Little to no evidence, moderate evidence, strong evidence, very strong?).*

(5pts) Question 3: Let's look at some other possible predictors of melanoma rate: Average temperature, Average age, and Proportion of white, non-hispanic residents. Create a simple linear model for each of these numeric predictors (i.e., create three simple linear models) and look at the output.

Are any of these three variables stronger predictors of melanoma rate than sun exposure? Explain (there are at least two good ways you could decide this, but provide at least one).

Create a scatterplot of the strongest predictor of these three. Again, use state abbreviations in place of points. Follow all formatting guidelines from question 1. **Include an image of this plot in your report.**

(5pts) Question 4: While it is possible to create a model with two numeric predictors, let's focus on the simpler case of using one numeric predictor with one binary predictor.

Using an `ifelse` statement, create a new variable in the dataset called `white_binary` that records "High" if the state's proportion of residents being "white, non-hispanic" is above 0.7 and "Low" if the proportion is below 0.7.

Include your ifelse statement code

Create a scatterplot with Sun exposure on the x axis, color the data points based on whether the proportion of white residents is above 0.7 or not, and put Melanoma rate on the y axis.

- Continue to use the state abbreviations in place of points.
- Add best fit lines for each group (i.e., visualize what an **interaction** model would look like).
- Add an appropriate title
- Adjust the axes labels. Also adjust the color legend to say "Proportion White." You can do that by entering a color = "…" argument into the labs function.

Include an image of your plot here *and* your code for this plot.

(5pts) Question 5: In the previous question, we plotted an interaction model (allowing for different slopes). Now, create a linear model with these same 2 predictors (the sun exposure variable, and white_binary) that includes an interaction term.

Copy or screenshot the summary output of your **interaction model** into the report.

Does the relationship between a state's average sun exposure and that state's melanoma rate appear to be dependent on the proportion of white, non-hispanic residents? *Justify your answer using your model summary. It also may help you to consult your graph from the previous question to see how this visual relates to the model output.*

(5pts) Question 6: Now, create an additive model for Sun exposure and white_binary as predictors of Melanoma rate, and consult this model to answer the following questions:

Copy or screenshot the summary output of your **additive model** into the report

Is there evidence that sun exposure has a linear relationship with melanoma rate while controlling for the proportion of white, non-hispanic residents in a state? *Briefly justify*

Contextually, how might the proportion of white, non-hispanic residents act as a *confounder* to the association between sun exposure and melanoma rate? *(Check back to Chapter 9 and look for the confounding diagram. Think about what that might look like here!).*

(5pts) Question 7: As a final challenge, let's create a map of the United States and fill color each state by their melanoma rate.

First, start by installing and librarying the package `maps`.

```
install.packages("maps")
library(maps)
```

Second, let's pull out a specific dataframe saved in `maps` that can help us outline the 48 contiguous U.S. States. We can do that with the following code. *If you get an error message that `map_data` couldn't be found, keep in mind that map_data is actually a tidyverse function (did you load it?)*

```
MainStates = map_data("state")
```

`MainStates` contains latitude and longitude coordinates that can be used to outline a map. Go ahead and click on it in your global environment to open it up and see what it looks like!

But before we can create our map, we'll need to merge in our melanoma rate data. To do that, we need a common column. Our `"Name"` column in Melanoma contains state names, but in MainStates, it's called `"region"`. So **third,** let's change that column title in the MainStates data frame to `"Name"`.

```
names(MainStates)[names(MainStates) == "region"] = "Name"
```

Now we're ready to merge! **Fourth**, run the following function to merge these two data frames. Go ahead and save this merged data frame under a different name (I suggest `Melanoma_Map`) so that we can still use the original data frame for other questions.

```
Melanoma_Map = inner_join(MainStates, Melanoma, by = "Name")
```

We're finally ready to create the map! Here are your instructions:

- In the aes line, map the latitude and longitude coordinates to the x and y dimensions (open up Melanoma_Map to see what they are called in the dataframe!)
- Also in the aes line, assign `group = group` and assign the melanoma rate variable to be represented as a fill color Use geom_polygon(), and set color = "black" inside this argument to insert black state borders

- Code in a fill color palette (check the Customizing ggplot2 tutorial!) and choose a custom color palette. The default palette is blues, but let's change it up. Choose something that makes clear where each color falls on a spectrum. Be careful _not_ to choose a diverging palette.
- Add the `theme_classic` to create a blank backdrop
- Add a title to your plot to clarify that this is the Melanoma Rate per 10,000

Insert your map graph

Which two states stand out as having the highest melanoma rates?