

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**
 - o 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)

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/m²)

lat: Latitude of the center point of the state

lon: Longitude of the center point of the state

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

(5pts) Question 1: Let's start by looking at the melanoma rates across U.S. states. By following these steps, we will create a map of the United States, and color each state by their melanoma rate. This should give us a little more insight to where melanoma rates are high.

Start by installing and librarying the package "maps" .

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

This next step will call up some U.S. state map border data—let's save it as a dataframe called "MainStates." Feel free to click on it in your global environment to take a look. *Be sure you have libraried tidyverse first!*

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

Next, we want to merge this data frame into our Melanoma data frame. But to do that, we need a common column. Our "Name" column in Melanoma contains state names, but in MainStates, it's called "region". Run the following command to change the MainStates data frame column to "Name". Then, we will be able to easily match up these columns when we merge the data frames.

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

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:

- Make a ggplot with **Melanoma_Map** as your data
- In the aes line, set `x = long`, `y = lat`, `group = group`, and `fill = mel`
 - o Note, these are variable names from *Melanoma_Map*, not from *Melanoma*
- Use `geom_polygon()`, and set `color = "black"` inside this argument to insert black state borders
- Use a fill color palette (with `distiller`) to choose a custom color palette. The default palette is blues, but let's take this opportunity to pick a palette with more dynamic differences on each side of the scale.
- Add a background theme that provides a blank background (check the "Customizing ggplot2 tutorial, or search online for one!)
- Add an appropriate title

Insert your map graph

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

Open the following short article about melanoma rates across U.S. States and read the first three paragraphs: <https://www.cancercenter.com/community/blog/2016/05/unhealthiest-states-for-skin-cancer-risk-may-surprise-you>

This article 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*.

Be sure to use the **original Melanoma data** you imported from this point on—the `Melanoma_Map` dataframe you created was only for Question 1.

(5pts) Question 2: Let's look at the original Melanoma data with one row for each state. Create a scatterplot using sun exposure as the predictor variable and melanoma rate as the response variable. However, *in place* of using `geom_point`, substitute in the following `geom` code:

```
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 a title
- Adjust the x and y axis titles to be fully written in 1-3 words (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 3: Now, create a simple linear model with sun exposure as a predictor of melanoma rate.

Copy or screenshot the summary output 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 4: 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 how you made that determination.

Create a scatterplot of the strongest predictor of these three. Again, use state abbreviations in place of points. Follow all formatting guidelines from question 2. **Include an image of this plot in your report.**

(5pts) Question 5: 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 6: In the previous question, we plotted an interaction model (allowing for different slopes). Now, create an 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 7: 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 if we already know whether the proportion white, non-hispanic residents in a state is above or below 0.7? 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!).