HW 2 – Math 58B

Your Name Here

due Thursday, Feb 2, 2023

Assignment Summary (Goals)

- calculating & interpreting correlations
- calculating & interpreting linear model

Note: you'll need many of the skills covered in lab 2 to complete the assignment! After Tuesday, the solutions to Lab 2 will be posted on Sakai.

Q1. LC Q Describe one thing you learned from someone in your learning commmunity this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Q2. Breaking Ice[1] Nenana is a small, interior Alaskan town that holds a famous competition to predict the exact moment that "spring arrives" every year. The arrival of spring is defined to be the moment when the ice on the Tanana River breaks, which is measured by a tripod erected on the ice with a trigger to an official clock. The minute at which the ice breaks has been recorded in every year since 1917. For example, the dates and times for the years 2000-2004 were:

2000	2001	2002	2003	2004
May 1, 10:47am	May 8, 1:00pm	May 7, 9:27pm	April 29, 6:22pm	April 24, 2:16pm

The data file NenanaIceBreak.txt contains all of the data since 1917. Scientists have examined these data for evidence of global warming, which would suggest that the ice break day should be tending to occur earlier as time goes on.

(a) Examine a scatterplot of the day in which the ice broke (`date` coded in column 7 with April 1 = 1) vs. `year`. Does it reveal any association between the two variables? In other words, is there any indication that the day on which spring begins is changing over time? Explain.

(n.b., Don't worry about the earlier columns coded with month and year. For this problem, the focus is on the number of days since April 1.)

```
# data available from a URL (not an R package)
ice <- read_delim("http://www.rossmanchance.com/iscam2/data/NenanaIceBreak.txt", "\t")
```

(b) Determine and report the regression line for predicting ice break day from year. Also calculate the correlation coefficient and the value of R^2. Comment on what these reveal, including an interpretation of the slope coefficient.

(c) Let's say that the conclusion is strongly that the slope is non-zero (as measured by the p-value, but we haven't learned that yet). Would you say that it reveals evidence of a **strong association** or **strong evidence** of an association? Explain.

[1] From ISCAM, HW 5.39

(d) Do the data suggest that one can make better predictions by taking year into account, rather than simply using the average of the ice break days? Explain.

(e) What date would the regression model predict for the ice break-up in the year 2005? What about 2020? Explain why you should regard these predictions cautiously.

Q3. Identify relationships, Part II., IMS Section 7.5 #4[2] For each of the six plots (see text), identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

Q4. Graduate degrees and salaries., IMS Section 7.5 #14[3] What would be the correlation between the annual salaries of people with and without a graduate degree at a company if for a certain type of position someone with a graduate degree always made

(a) $5,000 more than those without a graduate degree?
(b) 25% more than those without a graduate degree?
(c) 15% less than those without a graduate degree?

Q5. The Coast Starlight, regression. IMS Section 7.5 #21[4] Hint: look at the text for the scatterplot and also for how to compute the slope and intercept by hand. By hand computing will not be a regular thing, but it doesn't hurt to do it once so as to understand the computation.

The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes). The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time.
(b) Interpret the slope and the intercept in this context.
(c) Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.
(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.
(e) It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.
(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

```
praise()
```

```
## [1] "You are kryptonian!"
```

[2] https://openintro-ims.netlify.app/model-slr.html#chp7-exercises
[3] https://openintro-ims.netlify.app/model-slr.html#chp7-exercises
[4] https://openintro-ims.netlify.app/model-slr.html#chp7-exercises