Midterm 2023 Questions

Russell Almond

2023-10-16

For each question write a couple of sentences (or phrases) which (a) answer the question and (b) explain your reasoning for giving the answer.

Ex. 1 The relationship is mostly linear as the smooth curve in the scatterplot is close to a straight line.

Ex. 2. There is a strong relationship between and because $R^2=.70”.

# Part 1 – Hurricanes

These data come from the [Data and Story Library](https://dasl.datadescription.com/datafile/hurricanes-2015/?_sf_s=hurricane&_sfm_cases=4+59943)

Most weather models note at relationship between the barimetric pressure and the peak wind speeds. A secondary question is, as the average temperature rises, is that relationship changing.

The data and analyses are available in [Midterm 2023 Part I](Midterm2023-Part1.Rmd) or in <Midterm2023-Part1.pdf>.

## Q 1.1 Stability Across Years

The data include both historical data from 1850–1925 and more modern data from 1980–2015.

Is there any reason to believe that the average pressure or wind speeds is changing across time?

How about the variability?

## Q 1.2 Marginal Distributions

Describe the marginal distributions of the central pressure and peak wind speed variables, particularly, discuss the skewness and kurtosis.

## Q 1.3 Relationship with Category

What is the relationship between Category, peak wind speed, and central pressure?  
In particular, is Category adding new information or is it redundant with variables already in the model?

## Q 1.4 Linearity

Is the relationship between peak wind speed and central pressure linear?

## Q 1.5 Linear or Log Scale

Is it better to use peak wind speed or log (peak wind speed) as the variable?

## Q 1.6 Summary

Summarize the model you picked in the previous question. Include the equation of the line and either a -test or an -test.

## Q 1.7 Outliers

Looking at the dfbetas identifies two outliers. How strongly do these outlier affect the conclusions?

[For brevity sake, I randomly selected the model on the linear scale for this part, so use it regardless of how you used the previous two questions.]

## Q 1.8 Recent versus historical, same or different

There is a big gap in the data in the middle of the 20th C. Look at the two models fit with the hurricanes before and after that gap (as well as the scatterplot). Is there evidence to suggest that the recent (after 1980) data support a different model than the historical (before 1925) data?

## Q 1.9 Recent versus historical, Statistics

Is the difference in slopes bigger than we [meaning people who generally understand statistics] would expect if the only difference was random error?

## Q 1.10 Conclusions

What can we [same meaning] conclude about the relationship between central pressure and wind speed? Does this match the expectations from theory? Is the historical relationship different from the recent one?

#Part 2 – Depression and the Internet

These data come from the [Data and Story Library](https://dasl.datadescription.com/datafile/depression-and-the-internet/?_sf_s=depression&_sfm_cases=4+59943)

73 households were given free internet access in return for agreeing to being tracked. Several statistics were reported before and after.

Note that in 1998, home internet access was probably dial-up with broadband (cable/DSL) speeds only available to early adopters.  
Also, *Facebook* and other social media sites were just starting out.

The data and analyses are available in [Midterm 2023 Part I](Midterm2023-Part1.Rmd) or in <Midterm2023-Part1.pdf>.

## Q 2.1 Sample Relevance

Given the sample was taken in 1998, are the relationships discovered in these data likely to carry over to 2023?

## Q 2.2 Unit of analysis

The sample consists of 169 people from 79 households. Are the 169 observations independent?

## Q 2.3 Marginal Distributions

Characterize the marginal distributions of Internet Use and Change in Depression.

## Q 2.4 Teen vs Adult

Is the distribution for the two variables examined above different for teens and adults?

## Q 2.5 Linearity

Is the relationship between (Internet Use) and (Depression Change) mostly linear?

## Q 2.6 Linear versus Log

Would it be better to use Internet Use or log(Internet Use) for ?

## Q 2.7 Summary

Write a summary for the model you chose in the previous step. Include the equation of the line and a -test or an -test.

## Q 2.8 Age, Gender and Race

Are there differences in the model between Age (teen vs adult)? Gender (identifies male vs identifies female)? or Race (white vs non-white)?

## Q 2.9 Internet Vegans

There are 10 individuals with 0 internet use. Call them “Internet Vegans”.  
How sensitive are the model results to their inclusion.

## Q 2.10 Conclusions

What can we [people with some statistical training] conclude about the relationship between internet use and depression? In particular, does it support the news reports about the Kraut et al. (1998) study (from which the data come) which clearly concluded that internet use causes depression?