Homework 3

Sahana Sarangi

15 April 2024

Problem 1: Laverne, Gordon, and Tanya are trying to model the price of gas per gallon in their city. Each of them tries a slightly different approach.

- Laverne constructs a line of best fit for the data, and performs an F-test. She gets an R^2 of about 0.02 and an F-score of 2.04, and based on that she decides that we can't conclude anything with 90% confidence or better.
- Gordon takes Laverne's results to mean that we can't accurately model gas prices with just one line, so he tries a multipart linear model. He tries a whole bunch of different options, and eventually comes up with a five-part linear model which has an R^2 of 0.99, and an F-score of nearly 10000.
- Tanya recalls from the news that, at a certain point during the period of time they have data on, a toll was added to one of the major highways in the area. She hypothesizes that the toll may have changed people's driving habits and therefore changed the demand for gas, so she breaks up the data into "before the toll" and "after the toll" data sets, and constructs lines of best fit for each one, resulting in a two-part linear model. Her model has an R^2 of 0.76, and the F-scores of the pieces are 27 and 660, respectively.

Rank these three approaches from "most trustworthy" to "least trustworthy", and from "most successful" to "least successful". As always, explain your reasoning.

Solution: Gordon's approach is the least trustworthy because misinterprets Laverne's results, picks a five-part linear model for no particular reason, and got a suspiciously high F-score. Just because Laverne got a low R^2 and low F-score does not necessarily mean that we can't accurately model gas prices with just one line. Gordon also does not provide a strong case for why we should model gas prices using five parts instead of two, three, or four. Having an F-score of almost 10000 is also very suspicious, as even for a sample size of 3 with 99% confidence, the F-score is still only 4052.18.

Gordon's work is also the least successful. Because he has picked a model containing five individual line segments, he cannot simply "add on" to his model new gas prices that appear later. Laverne could because she has a one part linear model that can extend infinitely. Gordon's model is also a case of overfitting, or having the model be too perfect. Because he has incorporated potential errors in the data (hence the almost perfect R^2 score), even if he were able to use his model to model later gas prices, they could be somewhat inaccurate (more so than Tanya's or Laverne's models).

Laverne's work is the most trustworthy because she is the only one who recognizes the possibility for error within her model. Unlike Gordon and Tanya, she decides that nothing can be concluded with 90% confidence or better. While Tanya's work is somewhat successful because she actually provides reasoning for her decision to use a two-part model, she does not mention potential error.

Tanya's work is the most successful because she has a good reason for picking a two-part model (accounting for gas tolls) and has fairly high (but not suspiciously high) R^2 and F-scores. Tanya's model is also likely to be fairly good at modeling later gas prices as well, as she her "after the toll" segment could simply

continue on infinitely, or for however long makes sense. While Laverne's work has the potential to be the most successful because she accounts for error, her model still has fairly low R^2 and F-score values unlike Tanya.

Problem 2: In (a)-(e), a study or experiment is described. For each of these situations, decide whether or not bias was likely to have affected the results. If it was, describe the most significant source of bias.

(a) Dr. Carter is conducting a study on the health effects of a nutritional supplement called Vitamin Q. She was chosen to run the study because she is familiar with these supplements, since her business - Carter's Carp Capsules - sells them. Dr. Carter selects 1000 people at random from the lists of customers who have purchased products containing Vitamin Q (from any company in the United States), and interviews each one about their health after taking the supplement. Based on this data, Dr. Carter concludes with 95% confidence that people who take Vitamin Q experience increased mood and energy, and reduced incidence of illness, for up to three years after their last dose.

(b) Dr. Jackson wants to determine the popularity of Egyptology (the study of ancient Egypt) among Americans. He puts out a full-page ad in the back of Egyptology Today (the premier journal in the field) asking people to take his survey on the topic. Based on the data he collects, he finds that 97% of Americans report at least a "strong interest" in Egyptology, and 83% of Americans report that Egyptology is a "passion" of theirs.

(c) Dr. Teal'c is investigating whether wearing a mask is an effective way to limit the spread of COVID-19. To do that, he randomly selects 100 counties in the US (including major cities, small towns, and rural communities). For each of the counties, he computes the percentage increase in the total number of COVID-19 cases in that county for each day of the two years since the pandemic began. He then divides this data into two groups, depending on whether that county had a mask mandate during that time, and computes confidence intervals for the means of each group. Based on this calculation, he concludes with 95% confidence that the growth rate of COVID-19 cases actually increased during mask mandates.

(d) Dr. O'Neill is trying to determine the validity of astrology as a discipline. He's been told that people born under the sign of Mercury are more prone to injury, so he's collected large amounts of data on birth dates and hospitalizations going back fifty years. Unfortunately, the first few strategies he tries - confidence intervals, correlation, best-fit lines - are inconclusive, so he can't draw a conclusion one way or another. So he keeps trying, trying out different combinations of data and techniques, until he finally (on the 23rd attempt) finds one which allows him to conclude that Mercury does not have an impact on injury rates.

(e) Dr. Hammond is conducting a study on life-threatening accidents. He surveys a massive number of people, more than a million random subjects, about accidents they've been in and the medical ramifications. Among various other results, Dr. Hammond discovers that only 0.001% of accidents requiring hospitalization actually result in loss of life.

Part (a) Solution: Bias has affected the results. The bigget source of bias here is conflict of interest bias. Vitamin Q is one of Dr. Carter's products and is also the subject of this study, meaning that the results of this study are likely to have some impact on Dr. Carter and her business. She is likely to be biased towards gaining results that promote Vitamin Q rather than having results that make Vitamin Q look bad. This seems to have been the case, as the results of the study were in favor of Vitamin Q.

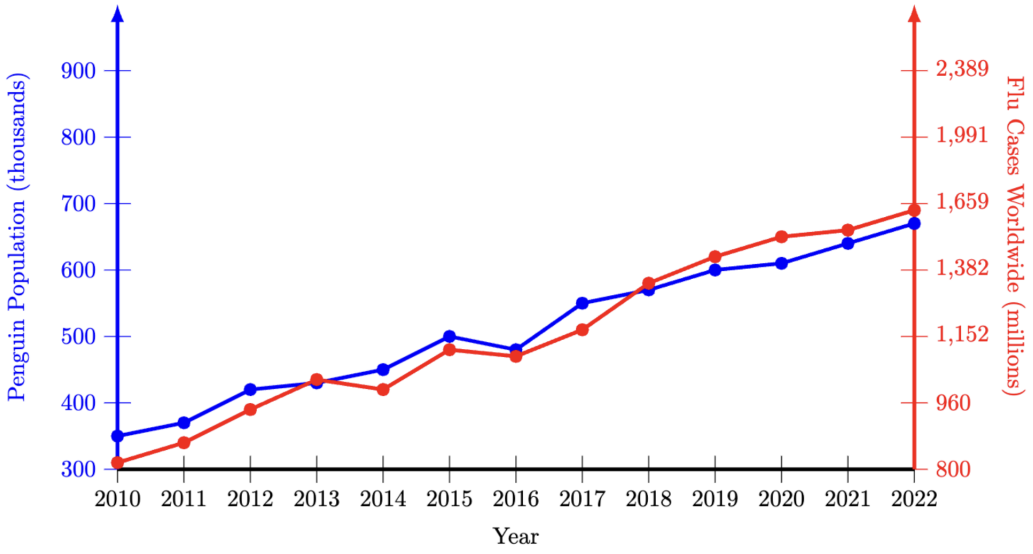
Part (b) Solution: This is a biased conclusion. The issue is that Jackson's sample set is not representative of Americans as a whole. He is collecting data from respondents who read Egyptology Today, which is a magazine that people would mostly only read if they were interested in Egyptology. So naturally, he finds that 97% of respondents have a strong interest in Egyptology and 83% describe it as a passion. If Jackson's sample set were to include people randomly drawn from the American population rather than readers of Egyptology Today, he would likely have very different results.

Part (c) Solution: While it seems like the conclusion was incorrect, the study itself is unbiased. There are no significant sources of bias that can be determined from the information we have.

Part (d) Solution: This conclusion is biased. the issue is that O'Neill is trying his best to look for some conclusion that is clearly not there. His first 22 attempts tell him that a conclusion cannot be drawn either way (which is the overwhelming majority of his attempts), but he decides that "inconclusive" is not satisfactory enough and uses his 23rd attempt as his final conclusion. This is biased because he is ignoring the first 22 "inconclusive" results that he got and is using only the 23rd as his final conclusion.

Part (e) Solution: This conclusion is biased. Obviously, very few of the respondents will report loss of life as all the people he surveyed must be alive. The sample here is not representative of people in accidents of all different medical ramifications, as it specifically excludes those who experienced loss of life. Therefore Hammond cannot make any conclusions on people who have experienced loss of life due to accidents, as his sample wouldn't include those people.

Problem 3: Professor Pippin has spent most of his adult life deeply suspicious of penguins. He's convinced that they're up to something. After much effort put into grant-writing, he's finally managed to secure funding for a study, and he comes up with data relating penguin population in Antarctica to the number of cases of the flu worldwide. He presents his data like this:



Based on the clear, close relationship between the two trends, Professor Pippin concludes that the Antarctic penguins are engaged in a decades-long scheme to weaken the human race through weaponized influenza. Professor Pippin's penguin paranoia aside, is his data visualization misleading? If so, explain how it should be improved.

Solution: The chart is misleading. Firstly, having two different vertical axes is very confusing. Pippin could have created two different graphs: one for penguin population vs. year and one for flu cases vs. year. Secondly, the penguin population is being measured in thousands while flu cases are being measured in millions. It is confusing to have 400000 penguins be at the same vertical level as 960000000 flu cases. Furthermore, neither the blue nor red axis starts at 0. This is one of the biggest red flags that a graph is misleading. On top of this, the red axis does not have consistent intervals or have any indicator that it intends to not have consistent intervals. The difference between 800 and 960 is not the same as the difference between 1152 and 960, but 800 and 1152 are equidistant from 960 on the axis. The intervals on the blue axis are also not consistent with the intervals on the red axis.

The blue axis and red axis together make this graph very misleading. Using two vertical axes on a graph is unconventional as it is and can be confusing. Instead, Pippin could have used separate graphs for penguin population and flu cases, made both vertical axes start at 0, and keep his intervals consistent in each graph (but both graphs don't necessarily need to have the same intervals).

Problem 4: Zebra pox is a very rare disease. It affects about one person in 100,000, but usually goes undiagnosed, because testing is difficult - until now! HealthCorp has come out with a home testing kit for zebra pox, available for the bargain price of $15, which HealthCorp claims accurately identifies whether a person has zebra pox in 99.9% of cases. More precisely, HealthCorp's published research shows that they have 99.9% confidence that this accuracy rate is between 99.875% and 99.925%. Is HealthCorp's zebra pox home testing kit a good test? Support your answer with detailed, concrete calculations or examples.

Solution: No, HealthCorp's home testing kit is not good. While an accuracy rate of 99.875% to 99.925% seems very good, it is actually not when we consider a group of 100000 people, for example. If there is an accuracy rate of 99.875% to 99.925%, this means the home testing kit is wrong about 0.075% to 0.125% of the time. This means that in a group of 100000 people, 75 to 125 people will get incorrect results. Among these 75 to 125 people, most will be false positives, but there is a chance that the 1 in 100000 person that has zebra pox will get a false negative. While 99.875% to 99.925% accuracy rates appears to be very good (especially with a solid confidence level), they are not actually good accuracy rates and will give a lot of people inaccurate results.

Problem 5: Consider this your last opportunity to do unethical statistics! For this problem, construct an argument which shows that higher temperatures are associated with increased salaries. You may not falsify data or do incorrect calculations, but you may cherry-pick, bias your sample, or present your results in a misleading way.

Solution: The data for temperature relating to salary has been presented visually in a misleading way: