Vedant Desai

Data Science for everyone

February 15, 2023

Homework 1

1. The following statements arise from various theories. In each case, correctly identify the independent variable(s) (there may be more than one) and dependent variable(s) (there may be more than one).

(a) Increases in minimum wage increase unemployment.

Ans: Increases in minimum wage – independent variable, increase in unemployment – dependent variable.

(b) Rainy weather decreases bicycle usage.

Ans: Rainy weather – independent variable, decrease in bicycle usage – dependent variable.

(c) Age, blood sugar levels, and sedentary lifestyle are all related to heart rate.

Ans: Age, blood sugar levels, and sedentary lifestyle – independent variable, heart rate – dependent variable

(d) In 1850s London, cholera incidence varied by distance to the Broad Street pump.

Ans: Distance to the broad pump – independent variable, cholera incidence – dependent variable.

(e) Heavier planes as well as faster ones consume more fuel.

Ans: Heavier planes and faster planes– independent variable, consume more fuel - dependent variable.

2. Associations between variables are straightforward to find—but need not be causal. As we have seen, confounders can cause problems. For each case below, give a plausible confounder for the potential causal relationship claimed. Briefly explain why it might be a confounder.

(a) Whenever a thermometer displays a temperature that is below 0 degrees Celsius the water in the bowl next to it freezes. The thermometer going below zero causes the water to freeze.

Ans: Plausible confounder could be lowering the temperature around the bowl. This could be a confounder because lowering the temperature can cause both the thermometer to display below zero temperatures, and the water to freeze.

(b) Many of the great rock guitarists of history played a Fender Stratocaster. Playing a Fender Stratocaster causes a guitar player to be a great rock guitarist.

Ans: Plausible confounder could be that the Fender Stratocaster is a common guitar for rock music, and this causes many great guitarists play the Fender Stratocaster, and vice versa.

(c) Patients who choose to undergo a potentially dangerous experimental treatment often fail to recover. The dangerous experimental treatment causes patients to not recover.

Ans: Plausible confounder could be the initial illness or the injury which caused the patient to undergo surgery/treatment. This can be a plausible confounder because this can be an initial problem which causes the latter things.

(d) Professional soccer players who score many goals are often able to execute many acrobatic tricks with the ball. Learning acrobatic soccer tricks makes you more likely to score a goal.

Ans: Professional soccer players are athletic which can cause them to score more goals and perform more acrobatic skills.

3. Read The New York Times article, 'On Landing Like a Cat: It Is a Fact" (NYT cats.pdf). Answer the questions that follow.

(a) What is the "unit of analysis" or sample of the cited study?

Ans: The unit of analysis or the sample of the cited study are the 132 cats.

(b) The article refers to a "surprising" causal relationship (paragraph 6). What are the treatment (cause) and outcome in this relationship?

Ans: The higher the fall, the greater the chances of survival.

(c) With respect to the causal relationship in the previous question, is the study cited by

the article a randomized trial or an observational study? How can you tell?

Ans: It is an observational study because the medical centers recorded the distance and observed the trends related to the fall.

(d) What "theory" do the authors of the study put forward to explain the causal relationship

above?

Ans: The authors of the study, put forward the notion that after a certain time cats reach terminal velocity and after that the velocity doesn't increase which allows the cat to get into flying squirrel like position which increases air resistance and helps to distribute the impact more evenly.

(e)Would it be feasible to test this theory by replicating the study with randomized treatment? What type of concerns would arise? Briefly explain.

Ans: Yes, it would be feasible by replicating the study to get a larger sample size and get more data on the cats that were not brought into the hospital. Although, there would be ethical concerns regarding the cats' lives.

(f) Indicate at least one possible unobserved confounder of the causal relationship identified

in the study. Explain how and why this could be a confounder.

Ans: A possible confounder could be more chances of getting stuck in a tree or objects which could slow down the impact. The smaller body of the cats could also be a factor in lowering the terminal velocity.

4. Now read the article, "Do cats always land unharmed on their feet, no matter how far they

fall?" (SD cats.pdf). Answer the questions that follow.

(a) The article identifies a flaw in the 1987 JAMVA study. Explain what the flaw is.

Ans: The article states that the cats that were not brought to hospital were not considered in the observational study (probably the ones that didn't survive)

(b) In class we identified this type of bias or error with a name, what is that name?

Ans: Survivorship Bias.

(c) How does this bias invalidate the "flying squirrel hypothesis"?

Ans: The bias invalidates the "flying squirrel hypothesis" as it does not allow the observation to consider the cats that did not survive the fall.

(d) In his discussion of this flaw, Dr. Garvey suggests an additional piece of evidence in

favor of the flying squirrel hypothesis. What is this piece of evidence?

Ans: He states that the cats landed on their chests which in effect supported the "flying squirrel hypothesis."

(e) Is this additional evidence enough to dismiss the claim that the study is flawed? Explain why or why not.

Ans: The additional evidence does suggest that the study was flawed because it neglected the cats that would not have survived the fall and as a result were not brought to the hospital.

5. In 1968, Charles Reep and Bernard Benjamin published a pioneering study on association

football (also known as soccer). Among other statistics, Reep and Benjamin reported that

approximately 80% of goals scored resulted from sequences of three passes or fewer. To

calculate this statistic, Reep and Benjamin took all the goals scored in their data and counted

the number of passes leading up to each goal. Some readers concluded that making fewer,

more direct passes will increase the probability of scoring a goal.

(a) Identify the independent variable and dependent variable.

Ans: Fewer passes- Independent variable, Scoring a goal – Dependent variable.

(b) What is wrong with the above conclusion? That is, why is it difficult to draw inferences

about causation here? Use the specific term from class and explain the problem.

Ans: The conclusion did not consider the times when fewer passes resulted in a missed goal. This is also known as survivorship bias. The problem is that this skews the data in favor of the scored goals.

(c) After reading the study, a coach instructs her team to focus on sequences of at most

three passes. After implementing this instruction, the team scores more goals. Does this

prove the theory that short passing sequences increase the probability of scoring goals?

Why or why not?

Ans: This does not necessarily validate the theory because the sample size is too small and could be considered an isolated incident.

(d) Another coach remains skeptical, asking "What if a team focused on sequences of five or more passes? Would the team score more goals?" What type of reasoning is this? Use the specific term from class.

Ans: This is called a testing hypothesis. We do not have enough evidence to suggest that team would score more goals. This is changing the independent variable and testing the outcome.

Data Science Homework 1

February 15, 2023

```
[2]:  #Q6
      import numpy as np
      from statistics import mean
      myarray = np.array([110000,75000,73000,70000,65000,62000])
      sum_myarray = myarray[0]+myarray[-1]
      avg_myarray = mean(myarray)
```

      75833

```
[43]: a = 1
      b= "2"
      print(a+b)
      #This does not run because you are trying to add two different data types
```

```
      ---------------------------------------------------------------------------
      TypeError                                 Traceback (most recent call last)
      <ipython-input-43-0beb8d008c0a> in <module>
            1 a = 1
            2 b= "2"
      ----> 3 print(a+b)
            4 #This does not run because you are trying to add two different data types

      TypeError: unsupported operand type(s) for +: 'int' and 'str'
```

```
[4]:  a = 1
      b = 2
      print(a+b)
```

      3

```
[8]:  my_list = [62000,60000]
      my_list.append(60000)
      arr = np.array(my_list)
```

      [62000, 60000, 60000]

```
[13]: #Q7
      def multiplier(a,b):
          ''' The function multiplies a and b and then cubes it '''
          ans = (a*b)
          return (ans)**3

      print(multiplier(2,3))
```

216

```
[23]: def falsehood(a,b):
          '''The function checks if the multiplier function returns a positive value␣
      ↪or a negative one'''
          if multiplier(a,b)>0:
              return True
          return False
      print(falsehood(2,3))
      print(falsehood(2,-3))
```

True
False

```
[30]: def falsehood_1(a,b):
          ''' This function only tests if the multipier gives a positive value'''
          if multiplier(a,b)>0:
              return True
          return None
      print(falsehood_1(2,3))
      print(falsehood_1(2,-3))
```

True
None

```
[42]: i = 5
      while i<=12:
          print(i)
          i+= 1.5
          if i>12:
              break
```

5
6.5
8.0
9.5
11.0

[ ]: