Problem Set 8

Statistics 104

Due April 30, 2020 at 11:59 pm

Problem set policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, be sure to include the code in your solution.*

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Problem 1.

The Unit 9 lecture introduced a case study about the *Challenger* shuttle accident. A report published by the commission responsible for investigating the accident determined that the cause of the accident was due to the failure of an O-ring seal in the solid rocket motor due to the unusual cold temperatures on the day of the launch (about 30.9 degrees Fahrenheit).

A risk analysis conducted prior to the launch had incorrectly concluded that "temperature data [are] not conclusive on predicting primary O-ring blowby." According to the commission, "A careful analysis of the flight history of the O-ring performance would have revealed the correlation of O-ring damage in low temperature."

In 1989, a more rigorous analysis of the data used in the pre-*Challenger* risk assessment was published in the *Journal of the American Statistical Association*.[1] These data are available in the `mcsm` package as the `challenger` dataset.

- `oring`: binary indicator for whether at least one O-ring failed, coded `1` for at least one failure, and `0` for no failures
- `temp`: temperature at the time of launch, measured in degrees Fahrenheit.

Fit a logistic regression model to predict O-ring failure from temperature at time of launch.

a) Write the model equation estimated from the data.

b) Does the intercept have a meaningful interpretation in the context of the data?

c) Interpret the slope coefficient. Assess whether temperature is significantly associated with O-ring failure.

d) Calculate the odds of O-ring failure for a launch on a day with temperature 70 degrees Fahrenheit. Is it more likely that O-ring failure will occur than not? Explain your answer.

e) Calculate the odds ratio of O-ring failure comparing a launch on a day with temperature 60 degrees Fahrenheit to a day with temperature 75 degrees Fahrenheit.

[1] SR Dalal, Fowlkes EB, and Hoadley B. Risk analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. *Journal of the American Statistical Association*. 84: 408 (1989).

f) According to the model, what is the predicted probability of O-ring failure for a launch when the temperature is 30.9 degrees Fahrenheit?

g) Based on the work done in the previous parts of this question, comment on whether there was reasonable evidence to postpone the *Challenger* launch. Be sure to reference relevant numerical details and to address one important caveat for the result in part f).

Problem 2.

An individual is said to be anemic (or have anemia) if they do not have enough healthy red blood cells to transport oxygen to tissues in the body. According to the US National Heart, Lung and Blood Institute, anemia affects approximately 3 million Americans and is the most common blood disorder in the United States. Anemia can have many causes, including internal bleeding or nutritional deficiencies. Anemia is diagnosed by measuring the level of hemoglobin present in blood; hemoglobin is the protein in red blood cells responsible for oxygen transport.

Loss of energy and easy fatigue are common symptoms of anemia. Individuals with mild anemia may not recognize their symptoms. Detection of anemia typically occurs if a physician notices a change in energy level and requests a blood test, or if a routine blood test indicates low hemoglobin level. Anemia can be particularly serious in infants and young children, causing impairments in mental, physical, and social development.

Anemia is considered a severe public health concern, especially in under-resourced parts of the world where individuals have limited access to healthcare. In this problem, you will examine data from a study examining the health status of 120 children living in a slum area of a large city in Southeast Asia. The aim of the study was to estimate the prevalence of anemia and investigate possible predictors of anemia. For children in the age range of this study, hemoglobin level less than 10.5 g/dL constitutes a diagnosis of anemia. For low-income children in this age range in the United States, the prevalence of anemia is approximately 15%.

The study collected demographic information in addition to measuring health variables. Family income level was recorded as wealth quintile within the sample; for example, a value of `1` for the `wealth` variable indicates that the child's family income level was in the lowest 20% for families with children participating in the study. Iron level was recorded via an assay in which negative values indicate iron deficiency and positive values indicate adequate iron level.

Data from the study are in the file `anemia.Rdata`. The following table provides a list of the variables in the dataset and their descriptions.

Variable	Description
`sex`	sex, coded `female` for female and `male` for male
`wealth`	relative measure of family income level
`diarrhea`	coded `Yes` for at least one episode of diarrhea in the past two weeks, and `No` otherwise
`whz`	standardized weight-for-height z-score relative to the national population
`age`	age in months
`iron`	blood iron level, in milligrams per kilogram (mg/kg)
`hb`	blood hemoglobin level, in grams per deciliter (g/dL)

Use the data to answer the following questions.

a) Explore the data.

   i. Describe the age and sex distribution of these children, with reference to appropriate numerical and graphical summaries.

   ii. Children are considered substantially underweight if, for their height, they are in the lower 5% of the distribution of weight. What percentage of children in this study sample are substantially underweight?

   iii. What proportion of children (in the study) are iron-deficient?

   iv. How does the prevalence of anemia in the sampled children compare with anemia prevalence in low-income children in the United States?

b) Calculate and interpret an appropriate measure of association between sex and presence of anemia.

c) Do the data support the claim that more than half of the female children in this slum area are anemic? Justify your answer.

d) Estimate the association between age and presence of anemia. Is an older child more or less likely to be anemic than a younger child? Justify your answer.

e) Fit a model to investigate whether the association between age and presence of anemia differs by sex. Interpret the model coefficients and summarize your findings.

f) Previous research has shown that iron deficiency is associated with anemia.

   i. Compare the odds of being anemic for a child whose iron level is -1.48 mg/kg to one whose iron level is 0.18 mg/kg.

   ii. Do these data provide evidence that iron level is significantly associated with the presence of anemia? Explain your answer.

   iii. Does the association between iron level and anemia persist after adjusting for the possible confounding variables sex, age, weight to height z-score, family wealth quintile, and recent episode of diarrhea? Explain your answer.

g) A public health official is about to hold a press conference announcing a decision to administer iron supplementation to all children living in the slum district, based on the results of the previous analysis. You have been asked to prepare the official for questions from the press. In no more than five sentences, provide answers to the following questions. Be sure to use language accessible to a general audience and fully explain your answers.

   i. What is the rationale behind the decision to provide iron supplementation to all children living in this district?

   ii. Does the study provide evidence that iron supplementation will reduce the prevalence of anemia in this population of children?

Problem 3.

Biological ornamentation refers to features that are primarily decorative, such as the elaborate tail feathers of a peacock. The evolution of ornamentation in males has been extensively researched; there are many studies exploring how male ornamentation functions as a signal of phenotypic and/or genetic quality to potential mates. In contrast, there are few studies investigating female ornamentation.

Some biologists have hypothesized that there is strong natural selection against overly conspicuous female ornaments. Bright or colorful plumage in females might be expected to increase the incidence of predation on nests for species in which females incubate eggs. Female ornamentation might also undergo positive selection, functioning in sexual signaling like male ornamentation, and indicating desirable qualities such as high immune function.

The data in the file `rubythroats.Rdata` are from a study of 83 female rubythroats, a bird species in which both males and females exhibit a brightly colored red patch on the throat and breast (referred to as a "bib"). In rubythroats, females incubate the eggs, while males provide food to females to facilitate uninterrupted incubation.

- `survival`: records whether the bird survived to return to the nesting site the subsequent year, `yes` if the female was observed and `no` if the female was not observed
- `weight`: weight of the bird, measured in grams
- `wing.length`: wing length of the bird, measured in millimeters
- `tarsus.length`: tarsus (i.e., leg) length of the bird, measured in millimeters
- `first.clutch.size`: number of eggs in the first clutch laid during the first year that the bird was observed
- `nestling.fate`: whether the nestlings from the first clutch survived to fledging (`Fledged`) or were lost to predation (`Predated`)
- `second.clutch`: whether the bird laid a second clutch during the first year that the bird was observed, recorded as `Yes` for laying a second clutch and `No` for otherwise
- `carotenoid.chroma`: a measure of the abundance of red carotenoid pigment in feathers, as measured from a sample of four feathers taken from the center of the bird's bib. Larger numbers indicate higher levels of pigment in the feathers and a more saturated red color.
- `bib.area`: the total area of the bird's bib, measured in millimeters squared
- `total.brightness`: a measure of bib brightness, calculated from spectrometer analyses. Larger numbers indicate a brighter red color.

You will be conducting an analysis of the results in order to investigate how bib attributes and other phenotypic characteristics of female birds are associated with measures of fitness.

a) Fit a model to predict nestling fate from female bib characteristics (carotenoid chroma, bib area, total brightness) and female body characteristics (weight, wing length, tarsus length). Identify the slope coefficients significant at $\alpha = 0.10$, and provide an interpretation of these coefficients in the context of the data.

b) Investigate the factors associated with whether a female lays a second clutch during the first year that she was observed.

   i. Is there evidence of a significant association between nestling fate and whether a female lays a second clutch? If so, report the direction of association.

   ii. Fit a model to predict whether a female lays a second clutch from nestling fate and bib characteristics. Identify the two predictors that are most statistically significantly associated with the response variable.

   iii. Fit a new model to predict whether a female lays a second clutch using the two predictors identified in part ii. and their interaction. Interpret the model coefficients in the context of the data.

c) Investigate the factors associated with whether a female survives to return to the nesting site the subsequent year.

   i. Fit a model to predict survival from bib characteristics, female body characteristics, first clutch size, and whether a second clutch was laid. Identify factors that are positively associated with survival for the observed birds.

   ii. Fit a new model with only the significant predictors from the previous model; let $\alpha = 0.10$. Comment on whether this model is preferable to the one fit in part i.

   For parts iii. and iv., use the better parsimonious model of the ones fit in parts i. and ii.

   iii. Compare the odds of survival for a female who laid 5 eggs in her first clutch to the odds of survival for a female who laid 3 eggs in her first clutch, if the females are physically identical and both laid a second clutch.

   iv. Suppose female A has bib area 350 mm^2, total brightness of 35, carotenoid chroma 0.90, tarsus length of 19.5 mm, wing length 51 mm, weighs 10.8 g, lays 4 eggs in her first clutch, and lays a second clutch. Female B has bib area 300 mm^2, total brightness of 20, carotenoid chroma 0.85, tarsus length of 19.0 mm, wing length 50 mm, weighs 10.9 g, lays 3 eggs in her first clutch, and lays a second clutch. Compare the odds of survival for females A and B.

d) Biological fitness refers to how successful an organism is at surviving and reproducing. Based on the results of your analysis, briefly discuss whether female ornamentation seems beneficial for fitness in this bird species. Limit your response to at most ten sentences. You do not need to reference specific numerical results/models from the analysis.

NOTE: For full credit on the problem set, complete *either* Problem 4 or Problem 5. Clearly indicate which problem you have chosen in your solutions.

Problem 4.

Polychlorinated biphenyls (PCBs) are a collection of synthetic compounds, called congeners, that are particularly toxic to fetuses and young children. Although PCBs are no longer produced in the United States, they are still found in the environment. Since human exposure to these PCBs is primarily through the consumption of fish, the Environmental Protection Agency (EPA) monitors the PCB levels in fish. Unfortunately, there are 209 different congeners and measuring all of them in a fish specimen is an expensive and time-consuming process.[2]

The file `pcb.Rdata` contains data collected from 69 fish specimens; each row represents data from a single specimen. The first 7 columns contain the amounts of specific PCBs measured in units of nanograms per kilogram (ng/kg): PCB138, PCB153, PCB180, PCB28, PCB52, PCB126, PCB118. The variable `pcb.total` indicates the total amount of PCBs detected.

If the total amount of PCBs in a specimen can be estimated well using only measurements of a few PCBs, the cost of PCB assays can be greatly reduced.

In your written report for the EPA, present a model for predicting the total amount of PCBs in a specimen from the levels of individual PCBs. Be sure to briefly explain the work done to develop the model, evaluate the strengths and weaknesses of the model, and discuss whether the model can be effectively used to reduce the costs of monitoring PCB levels in fish.

Limit your written report to at most 1 page and include all R code and output in the appendix.

Problem 5.

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume (L) of air expelled after one second of constant effort; higher values indicate better lung function. In 1980, FEV was measured on 654 children ages 3 through 19 in East Boston, Massachusetts, as a part of a larger study assessing change in pulmonary function over time in children.

The FEV measurements (`fev`) are in `lung_function.Rdata`, along with a record of smoking status (`smoke`), age measured in years (`age`), sex (`sex`), and height in inches (`height`).

Is smoking associated with decreased lung function in children?

Conduct an analysis of the East Boston data to examine whether there is evidence of an association between smoking and decreased lung function in children.

In your written report, be sure to describe the study participants, explain the analysis approach, summarize the results, and discuss your findings in the context of the research question.

Limit your written report to at most 1 page and include all R code and output in the appendix.

[2] Data from *Introduction to the Practice of Statistics*, 7^{th} ed.