Statistics 215B Assignment 1 Due: 5pm Mon 22 Jan 2024

Overview

This assignment has two goals: to exercise your skills in using R for data analysis, and to recall basic ideas from descriptive statistics, visualization, hypothesis testing, and multiple linear regression. Your job in this assignment is to investigate the connection between maternal smoking and infant health, using data. You will accomplish this by working through a guided analysis, detailed below. This case study is adapted from Chapter 10 of Nolan and Speed (2000), but the presentation here is self-contained.

Please read the entire assignment before you begin your work.

Maternal smoking and infant health

Nolan and Speed (2000) present the following quotation from the 1989 Report of the Surgeon General:

> . . . cigarette smoking seems to be a more significant determinant of birth weight than the mother's pre-pregnancy height, weight, parity, payment status, or history of previous pregnancy outcome, or the infant's sex. The reduction in birth-weight associated with maternal tobacco use seems to be a direct effect of smoking on fetal growth.
>
> Mothers who smoke also have increased rates of premature delivery.

("Parity" refers to whether or not a pregnant woman has previously given birth. "Payment status" has to do with the type of the mother's pre-natal health insurance.) We can isolate two claims:

1. Mothers who smoke deliver premature babies more often than mothers who do not.
2. Cigarette smoking has a stronger relationship to infant birth weight than several other relevant covariates.

At the risk of stating the obvious, premature delivery and small, underweight newborns are bad things. The first step in deciding whether maternal smoking *causes* these bad outcomes is to figure out whether maternal smoking is *associated* with them; the latter is the content of these claims.

You will study the claims in turn. The dataset forming the basis of your analysis is (a subset of) the Child Health and Development Studies (CHDS), a large survey on all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, California. On the course website is the file `babies.data`. It contains observations (rows) for 1236 live male births. The variables recorded for each birth are given in the following table:

Name	Description
bwt	Newborn weight (rounded to the nearest ounce)
gestation	Length of the pregnancy (days)
parity	Whether the baby is (1) or is not (0) the first-born
age	Age of the mother at conception (years)
height	Mother's height (inches)
weight	Mother's weight (pounds)
smoke	Whether the mother smokes (1) or not (0)

What to submit

Write a report which addresses your findings about the claims. Summarize each claim in your own words, as you understand it. For each claim, outline why your analysis of the data ought to be informative, explain the practical meaning of the possible analysis outcomes, report what outcome you obtained, and describe your conclusions. Some specific guidelines appear in subsequent sections of this document. Refer to figures and tables obtained from your R session whenever it seems helpful. Please remember to give every figure a title, axis labels with units, and (where appropriate) a legend. I strongly encourage you to install and use the R package `ggplot2` to make your figures—once you learn how to use it, many otherwise difficult graphical tasks become simple one-line commands.

The report should be long enough to convey what you understood about the content of the claims, and how strong a case is made for or against them by this data. The report should be no longer than that. The report should be written using LaTeX, and submitted in pdf format.

Your submission should include three files:

1. a file `assignment1.pdf` containing your report;
2. a file `assignment1.R` containing all the R commands you used for your analyses;
3. a file `assignment1-transcript.Rt` containing a transcript of an R session in which `assignment1.R` has been run without errors.

Please submit these materials through the course website before the due date.

Preparing the data

- Download the data file from the website and load it into R, as a data frame named `babies`.
- The variables gestation, age, height, weight, and smoke all have some missing values. The code for a missing value is not exactly the same across the variables. Figure out the missingness code for each variable, then replace all occurrences of the missingness code with R's missing value code, `NA`.
- Some of the variables in the dataset are actually categorical, but are coded numerically. Convert these variables from numeric vectors to factors in the `babies` data frame, with appropriately named levels. Confirm the conversion worked by inspecting a summary of the data frame.

- Look at a small number of other descriptive statistics or graphics that might be helpful in getting an initial feel for the data.

Analyzing claim 1: guidelines

Claim 1 states: mothers who smoke deliver premature babies more often than mothers who do not. A full-term pregnancy is defined by the medical community as lasting 40 weeks. A premature birth is defined as occurring prior to the 37th week of gestation.

1. Make one or more suitable graphical comparisons of the gestation distribution for smoking mothers to the gestation distribution of non-smoking mothers.

2. Add to the `babies` data frame a two-level factor variable indicating whether or not each baby was born prematurely. Use this factor and the factor `smoke` to carry out a relevant tabular comparison of distributions.

3. Make a figure which allows the comparison in the previous bullet point to be carried out visually.

4. Use the same table to carry out one or more hypothesis tests of the null hypothesis that smoking and non-smoking mothers have the same rate of premature delivery.

5. A related question is whether the overall average gestation time is shorter for smoking mothers, compared to non-smoking mothers. Conduct one or more appropriate hypothesis tests.

6. If there are other statistics, tables, figures, tests, or analyses that seem useful or important to you in assessing claim 1, produce them and report on them.

Analyzing claim 2: guidelines

Claim 2 states: Cigarette smoking has a stronger relationship to infant birth weight than several other relevant covariates. The only other covariates available in the data for us to check are parity, age, height, and weight.

1. Compare the difference in the average birth-weight between smoking and non-smoking mothers to the difference in the average birth-weight between first-borns and non-first-borns. Conduct suitable hypothesis tests to accompany the comparison.

2. Divide the mothers into "tall" (above median height in the data) and "short" (below median height in the data). Repeat the comparison of the previous bullet point for babies born to tall versus short women (rather than for first-borns versus non-first-borns).

3. Do the same again, for mothers who are "heavy" (above median weight) and "light" (below median weight).

4. Make a multi-panel figure which allows the comparisons of the previous three bullet points to be carried out visually for whole distributions, rather than averages. Put the y-axes across the panels in exactly the same range, to ease visual comparison.

5. Fit a multiple linear regression of birth-weight against height, weight, and parity (but not smoking status). Summarize and check the fit.

6. Fit a second regression like the previous bullet point, but including smoking status. Compare the two regression models informally and formally. Interpret the results of the comparison.

7. What are pros and cons of the multiple-regression approach, as compared to the univariate comparisons you carried out initially?

8. If there are other statistics, tables, figures, tests, or analyses that seem useful or important to you in assessing claim 3, produce them and report on them.

9. (EXTRA CREDIT) Use the plotting package `ggplot2` to produce a single multi-panel figure which does the following: for each bin created in a three-way classification by ⟨tall/short, heavy/light, parity⟩, visually compare the birth-weight distribution of smokers versus non-smokers. Create the figure using a **single** R expression that involves only `ggplot2` functions. What advantages does this comparison have over the linear regression approach?

References

Deborah Nolan and Terry Speed. *Stat Labs: Mathematical Statistics through Applications*. Springer Texts in Statistics. Springer, 2000.