Introduction to Bayesian Modeling

Data Analysis Project

Due 10 May 2024 by 11:59pm (before Saturday).

For this project, you will work alone or in small groups ($\sim$ 2 people) to understand, analyze, and report on a data set. Details about the data set, as well as instructions for your work, will be provided below. In addition to these details and instructions, I will provide four additional files to help you get started with your work:

`bcarotene.csv`

This file contains the data for you to use in this project. It is structured as a comma-delimited data file, with each observation having its own line and values for variables separated by commas.

`bcarotene_JAGS_analysis.r`

This is a starting R file which uses the packages rjags and coda to perform a Bayesian analysis of these data. This should be available soon, and certainly by the start of May.

A BUGS starter file and model file are also available upon request.

Background

An important piece of drug research has to do with pharmacokinetics, or PK. PK studies seek to understand how a substance is absorbed by, metabolized by, and eliminated from the body. Pithily, PK can be contrasted with pharmacodynamics (PD) as follows: PD is the study of how a drug affects the organism, whereas PK is the study of how the organism affects the drug.

One common topic for PK studies is how a drug builds up in an organism's bloodstream over time. For various drug dosages, concentrations of the drug are recorded in the organism's blood plasma. In some cases, the concentration of the drug in an organism's plasma can also affect the plasma concentrations of other substances as well.

In the present dataset, 46 volunteers were randomly assigned to receive one of five doses of beta-carotene (0, 15, 30, 45, or 60 mg/day) for up to 11 months in a double blind fashion. All patients were on placebo (untreated) for months 0, 1, 2, and 3, then treated of all following months—so your data will show a large spike in serum beta-carotene levels in the fourth month, when patients begin taking the drug. The specific aim of this study was to determine how different dose levels affected the serum beta-carotene levels over time. In addition to measuring the plasma concentrations of beta-carotene by dose, we were also interested in examining whether there was any effect of beta-carotene supplementation on vitamin E levels in the plasma. Because both beta-carotene and vitamin E are lipid soluble (that is, they are dissolved in fats rather than water), it might be possible that increased levels of beta-carotene might affect vitamin E levels in the blood.

Variables

The following variables are available for you to consider:

`PTID`

Patient ID number. There are 46 patients in this data set.

`MONTH`

Month of study, coded 0-15 based on how many months have passed since patient's entry into study. Months 0-3 are used as a baseline, with beta-carotene treatment beginning in month 4. Note that for some patients, less than 15 months of data are observed.

`BCAROT`

Plasma beta-carotene levels (μg/mL). This is recorded every month.

`VITE`

Plasma vitamin E levels (μg/mL). This is recorded every month.

`DOSE`

Dose of beta-carotene. Treatment begins in Month 4 and is constant until the end of the study. Dosage options are 0 (placebo), 15, 30, 45, and 60 mg/day.

`AGE`

Subject age at Month 0.

`MALE`

Indicator variable for maleness (i.e. 1 if patient is male, 0 if patient is not male).

`BMI`

Patient's body mass index at Month 0. BMI is defined as $\frac{\text{weight in kg}}{(\text{height in m})^2}$.

`CHOL`

Patient's serum cholesterol level (mg/dL) at Month 0.

`CAUC`

Patient's area under curve (AUC) for serum beta-carotene during treatment, normalized to number of treatment months. This functions as an average level of serum beta-carotene over the entire course of treatment (Months 4+).

`VAUC`

Patient's area under curve (AUC) for serum vitamin E during treatment, normalized to number of treatment months. This functions as an average level of serum vitamin E over the entire course of treatment (Months 4+).

Instructions

The primary aim of your analysis is to quantify the effect of beta-carotene treatment on serum beta-carotene levels over time.

In addition, there are two secondary aims for your analysis. (1) Quantify whether the effect of treatment on serum beta-carotene differs by age, gender, BMI, or cholesterol. (2) Quantify the effect of treatment on serum vitamin E levels over time, and determine if serum vitamin E levels are correlated with serum beta-carotene levels over time.

Your project is to analyze these data to best address the scientific questions of interest. Your final analysis should be presented in the form of a report. The main body of the report should *not* exceed 12 pages. However, you can place additional information (e.g. diagnostic plots) in an appendix if you feel they are necessary.

Your report should look like a formal report to a statistically naive researcher or an interested lay person. Because a statistical analysis seeks to answer a scientific question, you should organize your report in the customary manner for presenting scientific findings, which I will outline and explain below.

1. **Abstract:** Provide a concise description of the scientific question, the data used to answer it, and what you found. This section should highlight the most important parts of your work, and should be around a quarter to a third of a page in length. *Group Work: This section should be done collaboratively.*

2. **Introduction:** A brief explanation of the underlying scientific questions. *Group Work: Choose one group member to do this section, and identify them in the final report.*

   (a) *Background:* Begin with a description of the scientific motivation for the analysis. You may want to do some background reading on pharmacokinetics and beta-carotene to help you with this section. Use your own words to explain the scientific motivation—don't just copy what I've written here. Make sure you discuss any scientific details that factored into your decision-making about how to conduct the analysis.

   (b) *Questions of Interest:* Follow this with a statement of the scientific aims of the study. Again, use your own words to explain these aims. Explain which of these aims you were able to address with these data and your analysis. Highlight any discrepancies between the proposed aims and your actual work.

3. **Materials and Methods:** An explanation of the data set used in this project, as well as the methods used in your analyses. *Group Work: Choose one group member to do this section, and identify them in the final report.*

   (a) *Source of Data:* Describe the source and sampling method for the data, if known. Describe the variables that are available and how they relate to the scientific topic. If there are missing data, discuss them here. If there are potential measured or unmeasured confounding variables, discuss them here and discuss how they affect your analysis. This section should not be used for descriptive statistics—they will come in the results section. This section is for explaining your data in detail to the reader.

   (b) *Statistical Methods:* Describe the methods you use for your analysis, to the best of your ability. This should be a low-level technical summary of your methods, including citations or explanations for non-standard techniques. Explain the basic philosophy behind the analysis techniques you've chosen to use. You may want to describe the software you used as well. You do want to describe how you evaluated the appropriateness of your models.

4. **Results:** Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done, unless they provide insight into the scientific question. DO NOT INCLUDE RAW OUTPUT FROM STATISTICAL PROGRAMS. When possible, use words instead of cryptic variable names. Use forms of estimates that have some meaning to a statistically naive researcher (e.g. odds ratios rather than logistic regression parameters).

Group Work: Each group member should be responsible for at least one of the sub-sections below. Identify which group member was responsible for which sub-section in the final report.

(a) *Descriptive Statistics:* Begin the results section by discussing the (non-model-based) characteristics of your sample. Tables and figures are key tools in this section, but you should also provide explanations of your findings in text. If there are characteristics of the data that will cause technical difficulties in your analysis, discuss and explain them here. The idea is to make sure the reader understands the data you will model, as well as any relevant complications in that data.

(b) *Models:* Explain the major model or models used to address your scientific aims. Use tables and/or figures to present summaries of statistical inferences on those models (e.g. point estimates, probability intervals). In your text, explain how your models' findings address your scientific aims and provide practical examples of what your models say about relevant scientific quantities and relationships.

(c) *Model-Building / Model-Checking:* Discuss any statistical work you did to find and check your final model. If you considered a range of possible models in your analysis, you should explain how you picked the major models you reported on in the previous section. You should also include discussion of Bayesian convergence metrics and a sensitivity analysis examining the importance of your choice of prior for your final models.

5. **Discussion:** Discuss the conclusions which you feel can be drawn from the analyses. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses. *Group Work: This section should be done collaboratively.*

6. **Bibliography:** If you cited any works in your report, list them here. For this project, you will be principally interested in works explaining the scientific background for this study, and works explaining any methods you choose to use outside the scope of our textbook. Make sure to cite our textbook as well. *Group Work: This section should be done collaboratively.*

7. **Appendix:** If there is information you couldn't fit into the report that you still think is important for the reader's understanding, put it here. I make no promises that I'll look at appendices—but a well-constructed appendix is much more likely to get my attention than a number of pages of raw computer output. *Group Work: This section should be done collaboratively.*

The major theme of the above is to write to the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to anyone.

Also, lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can most often be summarized in a single sentence ("We found no evidence to suggest that the final model did not fit the data adequately.") You are reporting your major results and impressions of the data. If the reader wanted to see every detail, he/she would have to do the analysis himself/herself.