First Year Data Analysis Qualifying Exam
Department of Statistics, University of California, Irvine

Handed out: **Monday, June 26, 2017** Due: **Friday, June 30, 2017 at 6:00pm**

Turning In Your Exam: Email your complete solution to BOTH Rosemary Busta (rbusta@ics.uci.edu) and Dan Gillen (dgillen@uci.edu) by 6pm on Friday, June 30. LATE EXAMS WILL NOT BE ACCEPTED AND WILL NOT BE SCORED.

1 Background

In the early 1990s, published studies suggested that higher intake of beta carotene may be associated with a reduced risk of lung cancer. In particular, epidemiologic studies linked the intake of vegetables rich in beta carotene with a lower risk of cancer (especially lung cancer) and suggested that certain micronutrients are inhibitors of cancer. As such, beta-carotene was of particular interest as a preventative supplement for multiple cancers, and lung cancer in particular. However, before doing large scale clinical trials to investigate the role of beta-carotene on the risk of cancer, it was first important to understand the pharmacokinetics of beta-carotene supplementation. In this case, a small trial termed a *phase II trial* was conducted where researchers administered beta-carotene in various doses to volunteers and pertinent serum levels of beta-carotene were measured at regular intervals. Of particular interest in this type of study was quantification of how dose level affects the build up of beta-carotene in the serum over time, as well as how the dose level might affect other blood chemistries.

Forty-six 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 (Note: All patients were on placebo (untreated) for months 0, 1, 2, and 3, then randomized and treated at their randomization dose for all following months). The specific aim of the study was to determine how different dose levels affect serum beta-carotene levels over time. In addition to measuring the serum concentrations of beta-carotene by dose, researchers were also interested in examining whether there was any effect of beta-carotene supplementation on vitamin E levels in the serum (another biomarker that had been suggested as a possible cancer preventative agent). The rationale for this is that both beta-carotene and vitamin E are lipid soluble (that is, they are dissolved in fats rather than water), hence it might be possible that increased levels of beta-carotene might impact vitamin E levels in the blood. Other measured variables available in this data set include subject age, sex, weight, BMI, percent body fat, and serum cholesterol level at baseline.

Serum beta-carotene and vitamin E levels were measured at baseline (pre-treatment months 0, 1, 2, and 3), and at a number of follow-up visits after randomization (month 4 up to month 15). Note that not all patient completed all visits. When measured, these values were recorded along with the respective month of measurement. In addition, a time average (area under the curve) of both beta-carotene and vitamin E levels (during the period while on treatment) is also available.

2 Available Data

The data from the trial can be found at `"http://www.ics.uci.edu/~dgillen/FYE2017_DA/bcarotene.txt"`. Variables available in the data set include:

- `PTID` = patient ID
- `MONTH` = approximate study month(Note: All patients were on placebo (untreated) for months 0, 1, 2, and 3, then treated of all following months)
- `BCAROT` = serum beta carotene (ug/mL)
- `VITE` = serum vitamin E (ug/mL)
- `DOSE` = dose of beta-carotene
- `AGE` = subject age at randomization (yrs)
- `MALE` = indicator that subject is male
- `BMI` = subject body mass index (weight in kg / height in m^2)
- `CHOL` = subject serum cholesterol at randomization (mg/dL)
- `CAUC` = time average of serum beta-carotene while on treatment (area under curve)
- `VAUC` = time average of serum vitamin E treatment while on treatment (area under curve)

The first 10 lines of the dataset are given by:

```
> bcarotene[1:10,]
   ptid month bcarot vite dose age male   bmi chol   cauc   vauc
1     1     0    158 8.36   30  56    0 24.03  251 1100.4 9.0792
2     1     1    174 7.88   30  56    0 24.03  251 1100.4 9.0792
3     1     2    199 7.81   30  56    0 24.03  251 1100.4 9.0792
4     1     3    152 7.42   30  56    0 24.03  251 1100.4 9.0792
5     1     4   1095 9.46   30  56    0 24.03  251 1100.4 9.0792
6     1     5   1193 9.39   30  56    0 24.03  251 1100.4 9.0792
7     1     6   1228 9.97   30  56    0 24.03  251 1100.4 9.0792
8     1     6   2088 9.59   30  56    0 24.03  251 1100.4 9.0792
9     1     7   1248 9.87   30  56    0 24.03  251 1100.4 9.0792
10    1     8   1207 9.90   30  56    0 24.03  251 1100.4 9.0792
```

3 Scientific Goals

As previously noted the specific aim of the study was to determine how different dose levels effect serum beta-carotene levels over time. Researchers were also interested in examining whether there was any effect of beta-carotene supplementation on vitamin E levels in the serum. As such, you are asked to use the available data to address the following questions:

1. Is supplementation of beta-carotene associated with an increased time-average of serum beta-carotene levels? If so, is the effect of supplementation dose dependent?
2. Does beta-carotene supplementation affect the trajectory of beta-carotene levels in serum over time? If so, is the effect of supplementation dose dependent?

3. Does beta-carotene supplementation affect serum vitamin E levels over time? If so, is the effect of supplementation dose dependent?

4. Are serum vitamin E levels associated with serum beta-carotene levels over time?

4 General Instructions

You are to analyze the data to best address the scientific goals stated above. You should use appropriate and reasonably efficient statistical methods for estimating and quantifying uncertainty in associations. Your final analysis should be presented in the form of a brief report (no more than 10 pages including relevant tables and figures). You may place additional information (eg. relevant diagnostic plots) in an Appendix if you feel it necessary. The report should (at minimum) consist of the following sections:

1. Abstract - A brief summary of your basic findings.

2. Introduction - Background on the scientific problem, an introduction to the problem at hand, and what is to be addressed.

3. Statistical Methods - A clear discussion and justification of the methods you have used to analyze the data and the modeling strategy that you employed.

4. Results - A presentation of the results of your analysis that includes relevant and properly formatted tables and figures as well as complete and precise interpretations of your analytic findings.

5. Discussion - A synopsis of your findings, what you have achieved with respect to the scientific goals, any limitations your analysis may suffer from, and possible future directions to better achieve the scientific goals you set out to accomplish.

Your report should be well-written, succinct, and to the point! It should be written in a language that is understandable to the broad scientific community while precisely interpreting your finding. The discussion of statistical methods should be more technical than that provided to a non-statistical audience given the purpose of the report. It should be complete but brief - free of garbage and not-so-relevant material. It is critical that the appropriateness of your modeling choices be clearly justified in your report. You are encouraged to use relevant and well-formatted tables, plots and figures to help explain your findings. You may use any written references for this problem that you wish, **but you cannot communicate (talk, email, etc) with anyone about your analysis.**