36-707 Fall 2001, Project 1
Determinants of Plasma Retinol and Beta-Carotene Levels

1 Description of the problem

As a biostatistican employed at a research hospital, your main job is to support medical research conducted by members of the hospital staff, by helping with design and analysis of their studies. You recently attended a meeting in which one of the research MD's was seeking help analyzing some observational data that she and some colleagues have been collecting. Following is the text of a memo she sent to you summarizing the problem as she sees it.

> Previous observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. We designed a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. Study subjects ($n = 315$) were patients who had an elective surgical procedure during a three-year period, to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous.
>
> We are interested in knowing
>
> - Which personal characteristics lead to lower levels of plasma concentrations of the micronutrients retinol and beta-carotene?
> - How variable are plasma concentrations of these nutrients in humans? How are concetrations of the two micronutrients related to each other?
> - What other evidence is there in the data that might help us understand either (a) the relationships between various personal characteristics in this study; or (b) the physiological relationship between some personal characteristics and plasma concentrations of these micronutrients?
>
> Eventually we want to publish these data and analyses in a good medical journal, but also we hope that the results are strong enough that we can make a recommendation to the general public about how to improve levels of these micronutrients in their diets.
>
> Recall the graphs we showed in the meeting last week suggesting that alchohol levels are related to plasma retinol levels; that body mass (Quetelet index) and cholesterol tend to lower plasma levels of beta-carotene; while dietery intake, vitamin use and fiber intake seemed to raise them. We are interested in confirming all these relationships, as well as any other discoveries that we can defend with your analyses. We ran some quick regressions and ANOVA's in SAS last week and have discovered R^2's in the 40–50% range; it is probably quite simple to do much better.
>
> We know this won't take you very long; and we look forward to your analyses. We would continue this ourselves, but screening for a new study on the specific cancers related to low levels of each micronutrient gets underway next week and we just don't have the time.

A description of the data and the variables that have been collected can be found in Appendix A of this handout, and in the data area of the 707 webpages (see prdata.txt and prdata.dat). There are $k = 14$ variables and $n = 315$ observations.

Your report will be read by the MD herself; she intends to use the results you present as the basis of one or more publications if possible. She is typical of many research MD's: she expended a lot of effort at becoming an expert in medicine, and considers most other fields of study to be less difficult. She knows many words in statistics, since she reads and produces a lot of research studies herself, but doesn't have deep knowledge of the meanings and concepts behind the words. As a consequence she can be unintentionally misled if you use statistical jargon without adequate explanations; but she expects to be treated as a peer. As always, you also need to include technical appendices and references for the head of the biostatistics unit, a statistician who skims data analysis reports such as yours in order to monitor and improve the quality of the statistical consulting your group does.

While I am leaving it up to you to choose how and what to explore and analyze these data, the questions from the MD above should be helpful in getting started. I do not expect everyone to give in-depth answers to all the questions listed above: An excellent job on two or even one question is better than a mediocre job on everything. It would be nice, but not obligatory, for you to do a little EDA even on the questions you do not choose to answer in depth (however, you should make clear what you are or are not studying in your report).

Above all, feel free to use your imagination, other source materials, or whatever it takes, to delve into these data to highlight the questions that you think are important or interesting. The purpose of this project is to get you thinking about data analysis, and in the real world, you often won't have a strict set of defined goals.

2 Your mission

Your mission is to explore and analyze these data, and write up a report of your conclusions. Some issues to keep in mind include:

- *What is (are) the response variable(s)? What is (are) the predictor variables?* A variable or concept is named for each, but is it the right one? Is there a better way to measure each concept that the MD wants to learn about? Exploring different transformations and combinations of variables might leave you with more than 14 variables to work with.

- *Transformations.* Once you've settled on a response variable (or are looking at residual plots from a regression), it's worth thinking a little bit about whether it is approximately, or not at all, normally distributed. In the latter case, does transformation improve the picture?

- *Missing and unusual data.* In most real data sets like this there are missing observations and other irregularities. Often, it is difficult to know quite why the data are missing or what caused the irregularities. If you discover cases like these, what will you do with them? One idea is to drop the cases that have missing or unusual observations. Another idea is to try to correct them (fill in, or "impute", missing values; replace unusual values with something more sensible [how will you decide what is sensible?]).

- *Discrete and continuous predictors.* Some of the potential predictors in the data set are discrete; some are continuous. Thus, you may be fitting linear regressions, ANOVA models and ANCOVA models. Will the parameterizations ever be important to your interpretations of the results? You may (or may not) also find you need to create additional dummy variables.

- *Variable Selection and Validating your models.* We haven't discussed formal variable selection procedures yet in the course. In general you can get very far by choosing variables on intuition, and

narrowing or expanding the model according to what you see in F-tests, residual plots, and the like.

A simple way to validate a model chosen in this way, or to compare two models that are not nested, is to divide the data up into two pieces (if there are 1000 observations, `sample(1000,500)` would give you a random subset of 500 of them), do whatever model-fitting you like on the first half, and then look at the residuals of the second half, using the parameter estimates of the first half to obtain fitted values in the second half for each model you like. You can get fitted y's for new X values with the `predict()` function. Of course there's no reason that the two halves have to be equal, or even that you have to split the data up into only two pieces. The general idea is clear though: build your model on one part of the data, then use the other part of the data to assess the fit of your final model(s).

- *Unanswered Questions.* Is there anything more you need to know in order to interpret or extend the results of your work in the ways that the MD wants to do?

The above are a few technical and interpretational issues that will influence the kinds of analysis you do for your final report; there are other technical issues to think about too, like equality of variability of residuals throughout the range of data, normality of the residuals, existence of outliers, and all of the other usual things that you think about with linear models and ANOVA.

You will probably find that the methods presented so far in the course are adequate to answer these questions, and that's fine. Or you may feel you need to learn more; if you do, feel free to thumb through the textbook, ask me for advice, etc. At the other extreme, though the focus of the project is on linear regression, you may find that some questions of interest are so clearly answered by simpler exploratory and graphical analyses, that regression is not needed.

3 On Writeups

Two good on-line resources for writing reports are

- `http://www.rpi.edu/web/writingcenter`
- `http://filebox.vt.edu/eng/mech/writing/handbook/`
  [see also just `http://filebox.vt.edu/eng/mech/writing/`]

Both sites have links to writing centers at universities around the country, many of which in turn have pages that describe how to put together different types of reports. The type of report you are interested in is the "data analysis report" or the "lab report." Here is a generic outline for a data analysis report

1. Title Page: Title, your name and contact information, the date, course name/number;
2. Abstract: Informative summary of the whole report (100–300 words)[1].
3. Introduction: What is the work? Why is it important? What background is needed? How will the work be presented?
4. Description of data, how it was collected (if you know) or data sources (if you don't) and initial data summaries and EDA. This is a brief descriptive narrative, intended to get additional background and contextual information out to the reader and to acquaint the reader with the general features of the data. [Even readers who were involved in the data collection can learn something from this!]

[1] I will *not* also require an executive summary, which is a longer, even less technical, version of an abstract.

5. Analysis and results: Summary of in-depth analyses performed (what, why, how) and their results; This is a full descriptive narrative. Be complete, accurate, and precise, listing all the steps in a logical order. State what you actually did and what actually happened.
6. Discussion: Explain, analyze, interpret your results. Explain any errors, anomalies, or problems that occurred.
7. Conclusions: Draw conclusions from your results that are relevant to the person reading the report: answer the question "So what?" Constructive criticism of the study you are analyzing, and recommendations for improvement, are appropriate also.
8. Technical Appendices.
9. Bibliography and Credits.

Parts 3–7 constitute the meat of the paper for your primary audience. Usually, as with the fictional MD in this example, your audience is intelligent but unschooled in Statistics. So these parts should have as little technical material as you can possibly get away with: a few well-chosen graphs, the names of the techniques you applied and what parts of the data were used, and the minimum possible number of parameter estimates or p-values is the most you should think about. It is appropriate, and even recommended, to refer the reader to the appendix in Part 8 if you need to provide a more technical explanation for something.

Part 8 is for your secondary audience—me, or the fictional head of the biostatistics group in this example—and should follow closely enough the "story" of parts 5 and 6 that it is easy for me to see what technical material backs up which results and discussion. However, it should be possible for the "manager" to understand what you've done from parts 3–7, without looking at part 8 at all.

If I were doing this, I would write part 8 first, as a sort of edited and annotated "diary" of my work as I learn about the problem (maybe not too different from the handouts I prepare for class). Then I would write parts 4, 5, 6, using the highlights of my "diary", and finally I would write parts 7, 3, and 2, summarizing my work. I expect part 2 would be short, since you don't know any more about the problem than I've told you. It is not necessary to number these parts 1–9; you may also merge some parts if it seems natural to do so.

Please don't discuss this project with each other, although you may discuss it with me. I would like you to provide a list of any outside sources you consulted (books, Web pages, etc.) in Part 9 of your report.

Your grade on this data analysis report will be very subjective, and will depend on you selecting and adhering to a logical and readable format for the report; on the balance of inventiveness and appropriateness in the selection of methods and conclusions in your report; on the correct use of whatever exploratory, graphical, and/or fitting technique you use; and on the readability and understandability of the report when technical material is deleted.

4 Timeline, Deadlines

I would like to discuss with you your progress after about two weeks, and I would like you to hand in your final reports one week later. So the deadlines are:

- *Tue. Oct 23.* I will check a draft of your paper, probably during class time. Write as much as you can, but leave holes or bulleted "to do lists" in the parts that depend on analyses you haven't done yet or that aren't written yet (so I can see where you're going).
- *Tue. Oct 30.* Bring your final reports to class.

A Appendix: The Data

This datafile contains 315 observations on 14 variables.

```
Variable Names in order from left to right:
        AGE: Age (years)
        SEX: Sex (1=Male, 2=Female).
        SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
        QUETELET: Quetelet index (weight/(heightˆ2)); values above 27 kg/mˆ2
                  (female) or 28 kg/mˆ2 (male) indicate obesity
        VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
        CALORIES: Number of calories consumed per day.
        FAT: Grams of fat consumed per day.
        FIBER: Grams of fiber consumed per day.
        ALCOHOL: Number of alcoholic drinks consumed per week.
        CHOLESTEROL: Cholesterol consumed (mg per day).
        BETADIET: Dietary beta-carotene consumed (mcg per day).
        RETDIET: Dietary retinol consumed (mcg per day)
        BETAPLASMA: Plasma beta-carotene (ng/ml)
        RETPLASMA: Plasma Retinol (ng/ml)
```

Here are the first several lines of the data file:

```
   age sex smokestat quetelet vituse calories  fat fiber alcohol cholesterol
 1  64   2         2 21.48380      1   1298.8 57.0   6.3     0.0       170.3
 2  76   2         1 23.87631      1   1032.5 50.1  15.8     0.0        75.8
 3  38   2         2 20.01080      2   2372.3 83.6  19.1    14.1       257.9
 4  40   2         2 25.14062      3   2449.5 97.5  26.5     0.5       332.6
 5  72   2         1 20.98504      1   1952.1 82.6  16.2     0.0       170.8
 6  40   2         2 27.52136      3   1366.9 56.0   9.6     1.3       154.6
 7  65   2         1 22.01154      2   2213.9 52.0  28.7     0.0       255.1
 8  58   2         1 28.75702      1   1595.6 63.4  10.9     0.0       214.1
 9  35   2         1 23.07662      3   1800.5 57.8  20.3     0.6       233.6
10  55   2         2 34.96995      3   1263.6 39.6  15.5     0.0       171.9

   betadiet retdiet betaplasma retplasma
 1     1945     890        200       915
 2     2653     451        124       727
 3     6321     660        328       721
 4     1061     864        153       615
 5     2863    1209         92       799
 6     1729    1439        148       654
 7     5371     802        258       834
 8      823    2571         64       825
 9     2895     944        218       517
10     3307     493         81       562
```