Exam 1: Diabetes

36-402, Spring 2012

Due at 10:30 am on Tuesday, 6 March 2012

Instruction

Please read the problem background carefully, before beginning the data analysis.

There are multiple ways to do most of the problems, some of which work better than others. You will be graded not just on the technical correctness of your results, but also on the soundness of the reasoning you use to get to the results, and the clarity with which you communicate.

You have one week to work on this exam, and can use your notes, the textbooks, and indeed anything you find in the library or online, if that is properly acknowledged. However, **all your work must be your own**. You cannot work with classmates, friends, a tutor, or anyone else. If you are unclear about what is allowed and what is not, please check the university policy on cheating and plagiarism (`http://www.cmu.edu/policies/documents/Cheating.html`), or ask the professor.

Please include the following text in your write-up:

> I, YOUR NAME, have completed this examination honestly, without giving prohibited assistance to anyone, or receiving it from anyone.

If, for reasons of conscience, you are unable to make such an affirmation, let the professor know at once, to arrange for an oral mid-term.

Background

Diabetes is a family of diseases where the body does not metabolize sugar properly. Ordinarily, metabolism of blood sugar is regulated by a protein called insulin. In type I, the body stops producing enough insulin; in type II diabetes, the body stops responding to insulin[1]. While they are related, the treatments are very different: type I diabetics need regular injections of insulin, and benefit from treatments which increase the body's production of insulin; these are of no use for type II diabetes.[2]

C-peptide is a protein which is produced along with insulin, but is easier to measure, and provides an accurate proxy for their blood insulin levels. Our data set contains information about 43 patients with type I diabetes: their age when they were first diagnosed with diabetes, the logarithm of their C-peptide concentration (in picomoles per milliliter), and `base.deficit`, a measure of how acid their blood is compared to normal controls[3]. The questions of immediate interest are about predicting c-peptide levels from age and base-deficit. The questions of ultimate interest are about what can be done to increase insulin levels for patients suffering from type I diabetes.

[1] Type I and type II diabetes used to be called "juvenile" and "adult-onset", respectively. While it is true that most people who become diabetic as children have type I, while type II tends to develop later in life, there are plenty exceptions in both directions. (I know someone whose type I diabetes first manifested in his mid-30s, on his honeymoon.

[2] If you need to know more, there are much better sources than Wikipedia, such as the National Diabetes Information Clearninghouse, run by the National Institutes of Health.

[3] There are several reasons why this was of interest. One is that diabetes can lead to a metabolic condition called *ketoacidosis*, when the body, starved for sugars, starts breaking down proteins, with waste-products that make the blood acid (and give the breath a sweet, fruity smell); this can also occur with starvation, or extreme alcoholism.

Problems

1. (10 total) Fit an additive model,

   $$\hat{y}_{AM}(\texttt{age}, \texttt{base.deficit}) = \alpha + f_1(\texttt{age}) + f_2(\texttt{base.deficit})$$

   with c-peptide level as the response.

   (a) (5) Plot and describe your estimates of the partial response functions f_1 and f_2.

   (b) (5) Plot the predicted surface $\hat{y}_{AM}(\texttt{age}, \texttt{base.deficit})$. (Contour, heatmap and wireframe plots are all acceptable. Make sure the axes are labeled with variable names as well as numerical units.)

2. (15 total) *Conditional predictions*

   (a) (10) Plot the predicted c-peptide level for a five-year-old patient against `base.deficit` level, letting `base.deficit` run over the whole observed range, i.e., plot

   $$\hat{y}_{AM}(5, \texttt{base.deficit})$$

   as a function of `base.deficit`.

   (b) (2) Add to this plot two more lines, showing the predicted c-peptide level as a function of base-deficit for patients aged 10 and 12 years.

   (c) (3) Are the three lines for the three ages parallel to each other? Should they be?

3. (10 total) *Kernel model*

   (a) (5) Fit a kernel regression jointly to `age` and `base.deficit`. Plot the predicted surface $\hat{y}_{KM}(\texttt{age}, \texttt{base.deficit})$. Describe the surface and compare it to the surface for the additive model.

   (b) (5) Plot predicted c-peptide levels as functions of base-deficit for patients aged 5, 10 and 12 years, as in Problem 2. Are the three lines for the three different ages parallel to each other? Should they be?

4. (25 total) *Model comparison* Should we use a strictly additive model here, or should we allow for an interaction between `age` and `base.deficit`?

   (a) (8) Describe a procedure which could be used to decide between these options.

   (b) (4) Explain in what sense the model picked by this procedure ought to be better than the one it doesn't pick.

   (c) (5) Explain why this procedure *reliably* picks the better model.

   (d) (8) Apply the procedure to the data and report the result.

5. (25 total) *Predicting response to changes*

   (a) (5) On average, by how much would increasing the age of each patient by 1 year change their c-peptide levels? Answer in terms of the model you picked in Problem 4, and assume patients' `base.deficit` levels remain unchanged.

   (b) (5) On average, by how much would increasing `base.deficit` by one tenth of a standard deviation (i.e., moving it towards zero) change the c-peptide level? Again, answer in terms of the model you picked in Problem 4, and assume that patients' ages remain unchanged.

   (c) (5) Calculate standard errors for both of these average predicted changes. Be sure to explain both the method you used to find the standard errors, and why that method is appropriate to this problem.

   (d) (5) Calculate a standard error for the *difference* in these average predicted changes.

   (e) (5) If these two changes are equally easy to bring about, which one would be better? (Assume that higher c-peptide levels are better, generally, than lower ones.) How sure should you be of this answer? What would it mean to increase the patients' ages?

6. (15) Summarize your analysis in one page of prose. Refer to your work on earlier problems for details, but try, as much as you can, to be clear to someone who had not taken 402, or even 401.