Applied Nonparametric and Modern Statistics

Official title: Advanced Generalized Linear Models IV 140.753-754

Rafael A. Irizarry
Department of Biostatistics
Johns Hopkins University

Fourth Term 2001

Chapter 1

Introduction

A common problem in applied statistics is that one has an independent variable or outcome Y and various dependent variable or covariates $X_1, \ldots, X_p$. One usually observes these variables for various "subjects".

One may be interested in various things: What effects do the covariates have on the outcome? How well can we describe these effects? Can we predict the outcome using the covariates?, etc..

Statisticians usually assume that Y and the X's are random variables. Then one can summarize the above question by asking: what is $\mathrm{E}[Y|X_1, \ldots, X_p]$? We usually call $f(X_1, \ldots, X_p) = \mathrm{E}[Y|X_1, \ldots, X_p]$ the *regression function*.

It should be noted that for some designed experiments it does not make sense to assume the X are random variables. In this case we usually assume we have "design points" $x_{1i}, \ldots, x_{pi}, i = 1, \ldots, n$ and non-IID observations $Y_1, \ldots, Y_n$ for each design point. In most cases, the theory for both these cases is very similar if not the same. These are called the *random design model* and *fixed design model* respectively.

How do we learn about $E[Y|X_1, \ldots, X_p]$?

A common procedure is linear regression. One typically assumes

$$\mathrm{E}[Y|X_1,\ldots,X_p] = \sum_{j=1}^{p} X_j\beta_j.$$

Assuming that the conditional probability of Y is normal is quite common. However, if the range of this expectation is not continuous one can generalize to:

$$g(\mathrm{E}[Y|X_1,\ldots,X_p]) = \sum_{j=1}^{p} X_j\beta_j$$

with g a link function with real-valued range. It is typical to assume the conditional distribution of Y is part of an exponential family, e.g. binomial, Poisson, gamma, etc.... Many times the link function is chosen for mathematical convenience.

These model have the convenience that the parameters β usually have direct interpretation with scientific meaning. For example $\beta = 5$ may mean that a man that is one inch higher than another is expected to weigh 5 more pounds.

Another advantage is that, once an appropriate model is in place, the estimates have many desirable properties.

A drawback of these models is that they are quite restrictive. Linearity and additivity are two very strong assumptions. This may have practical consequences. For example, by assuming linearity one may never notice that a covariate has an effect that increases and then decreases. We will see various example of this in class.

In this class we will

- Start by introducing various smoothers useful for smoothing scatter plots $\{(X_i, Y_i), i = 1,\ldots,n\}$ where both X and Y are continuous variables.
- Set down precise models and outline the proofs of asymptotic results.
- Introduce local regression (loess).

- Examine spline models and some of the theory behind splines.
- Some smoothers are more flexible than others. However with flexibility comes variance. We will talk about the bias-variance trade-off and how one can use resampling methods to estimate bias and variance.
- After explaining all these smoothers we will make a connection between them. We will also make connections to other statistical procedures.
- We will examine the case were one has many covariates. One can relax the linearity assumption, assume additivity and use additive models. One can also forget the additivity assumption and use regression trees.
- After all this we will be ready to consider the case where Y is not necessarily continuous. We will generalize to this case and look at Generalized Additive Models and Local Likelihood.
- While examining all these subjects we will be considering various models for one data set. We will briefly discuss techniques that can be used to aid in the choice of such models.
- Finally we will look at a brief introduction of times series analysis.

By relaxing assumptions we loose some of the nice properties of estimates. There is an on going debate about specification vs. estimation.

We will begin the class talking about the case were the regression function f will depend on a single, real-valued predictor X ranging over some possibly infinite interval of the real line, $I \subset \mathbb{R}$. Therefore, the (mean) dependence of Y on X is given by

$$f(x) = \mathrm{E}[Y|X = x], x \in I \subset \mathbb{R}. \tag{1.1}$$

Sometimes, the need to estimate f arises when investigators have to decide among various explanations for a physical phenomenon, and existing subject-knowledge or scientific theory says nothing about f. In this case we collect data to see what it says. Exhibiting some aspect of f may then imply the confirmation or revision of a given theory.

The data to support such investigations are typically a set of n paired observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. These can be either a random sample of the joint distribution of $(\mathbf{X}, Y)$ as is the case for observational studies, or fixed input values $\{\mathbf{x}_i\}$, arising perhaps from a designed experiment.

So once we have the data what do we do?

If we are going to "model" (1.1), we gain insight into the important features of the relationship between Y and $\mathbf{X}$ by entertaining various descriptions of or models for f. Through this exercise we might:

- identify the width and height of peaks
- explore the overall shape of f in some neighborhood
- find areas of sharp increase or regions exhibiting little curvature.

The first three chapters of the class deal with this problem. We will then move on to the case where we have many covariates, then cases where the expectation needs to be transformed, and various other generalization.

Through out all these subjects we will be talking about finite sample theory, asymptotics, practical aspects, and computational consideration.