Modeling and Visualization

1. Introduction

The purpose of these notes is to discuss a few broad facets of statistical modeling and augment the class notes on classical regression diagnostics.

Models can be used for different purposes. In many practical situations it suffices for models work well and we may not care how they work as long as the do work. In many cases a black box model will suffice. However in scientific contexts, we want to learn and understand. We want to make educated decisions, make inferences rationally, and making predict to new circumstances based on understanding of how nature, people, societies work. We learn about the world through data and models. George Box said, "Essentially, all models are wrong, but some are useful." Some are useful for learning.

In Section 2 that I attempt to convey a statistical perspective on modeling. I wish might student could read the book "A Lady Tasting Tea" to get a better understanding of what I am struggling to communicate and to appreciate the history of contributions to statistical methodology and modeling

This learning has continued. An excellent reference on modern modeling is *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman. Section 2.3 of their book addresses linear regression and nearest neighbor methods. Section 3 addressed this topic trying to express in a broader, less specific context. I wish I had time to teach more Form the book

I don't see a sharp line between statistical and scientific visualization, and quantitative models. In *The Power of Maps*, Denis Wood says there is an agenda behind every map. I think there are models behind every graph: models of scientific knowledge, of data, of quantitative modeling, of production technology, of communication to audiences, and of human perception/cognition. I wish the class were to devote more time to quantitative modeling and the associated graphics.

2. Statistical models

Statistical models presume an underlying framework or process playing out over space time, or both. Observed data reflected the process in action. The goal is to better understand the process. A purpose of statistical models is to provide connections across space, time, and attributes that build scientific knowledge.

Why use statistical models? Consider health scientists studying changes in breast cancer mortality rates over time or rate disparities for different populations of people. Mortality rates usually have numerators and denominators. In looking backward don't we know who died from breast cancer and who was at risk of dying? The answer is that we have only partial knowledge. Sources of error for numerators include unreported deaths, incorrectly reported cause of data, attribute errors in patient records such as age and sex,

and transcription errors. Source of error for denominators relate to decennial census currency accuracy and to change births, deaths, immigration and migration. Low estimates of the population can lead to the appearance of high rates that cause concern. A reason for using statistical models is to reflect uncertainty that is related to the measurement process that provides data for our calculations.

The situation gets more and more complicate as the populations we seek to characterize become more specific. Cancer is most a disease of the older people. Should not the population being described be age specific? What about sex and race differences. Note that some men do get breast cancer. Not also that prostate cancer rates would look a lot lower if women were included. More detailed analysis could include all genetic and cultural differences and all environmental risk factors. In general there is much data that we don't have that is influencing outcomes.

We use statistical models in part because they provide structure-residual (deterministic-stochastic) decompositions that help us see patterns in the presence of noise and help us to produced predictions with bounds on the uncertainty. Often we think of predictions as relating to the future, but they can relate to locations, and to other populations as defined by attributes.

Part of the art of statistical modeling is in making choices about
1) the structure-residual (deterministic-stochastic) decompositions,
2) how much to elaborate the deterministic part
3) how much to elaborate the stochastic part

Part of the art of statistical modeling is related to data
1) designing experiments to obtain data that can answer the next questions in the progression knowledge
2) obtaining additional data to make better use of available data
3) transforming data to produce variables better related to the phenomena of interest

3. Model Evolution

The notion of model evolution has been well expressed by George Box and others. The following provides a narrow example to illustrate idea assessing model inadequacies. In the bigger picture, model refinement can also involve the strategic collection of data need to address questions raised.

In one view classical linear regression assumes model residuals are independent and identically distribution normal random variables with mean 0 and standard deviation sigma. The analyst often seeks to account for an increasing fraction of the variability in the data building on past experience that include previous modeling and accumulated scientific knowledge.

In modeling mortality rates early models often assumed that deaths followed a Poisson (or sometimes a binomial distribution.) The Poisson distribution has a since parameter that determines both the mean and variance. The parameter can be re-expressed as a rate per 100000 people per year times the number people in 100000's for each region. Experience quickly reveals there is additional variation (also called over dispersion and sometimes extra Poisson variation that this model does not capture. Having a single parameter for both the mean and variance doesn't work vary well in modeling lots of different mortality data. Thus the model needs to be changed fundamentally or adapted.

Many classic statistical models assume the presence of independent errors with mean zero and standard deviation sigma. Often the error distribution is assumed to follow the normal distribution. The errors are approximated by models residuals. Model criticism makes use of diagnostics that often involve residuals to help in assessing such assumptions. Classic issues related to problems related to correlated errors and to thick tailed error distributions. George Box used to talk about correlation as being the tiger loose in the room. Today there are books writing on time series and spatial models that address correlation. Today a host of semi-parametric and non-parametric methods are available to help deal thick tailed error distribution.

The detective work in model building can take various forms that are not little discussed. For example classical regression model estimates typically include s, an estimate of sigma. External knowledge may suggest alternative sources of variation related to sigma. If this knowledge is built into a model, but the resulting estimate of s ends up being 1.1 instead of 1.0, this is a clue that knowledge is not complete or there is some other problem. While the situation is a bit different due to the use of Poisson model rather that the normal model, the model evolution indicated above follows the pattern. The variation left after using the Poisson model was too big. There had to be something more.

Modeling approach continues to evolve. This is driven by computational power, a better of understanding of models, new kinds of data, and increase amounts and flow rates of data. Recent methods that are drawing my attention include

1) Lasso regression, a more modestly greed approach to forward stepwise regression.
2) Semi-supervised principal components developed to deal with such large numbers of variable that even lasso regression has problems
3) Rule Fit for supervise learning that can combine more traditional regression with tree building methods and deals with a large number of variable using lasso regression.

3. Data as the model

Every so often researchers discovered that they can model data almost perfectly. Unfortunately experience has shown that such models often don't perform very well on validation data that had been excluded from the modeling process or on new data. Tightly fitting the data is often a problem because there are so often idiosyncratic facets of the data that do not generalize. (The point of have error in the statistical models is to address to idiosyncratic facets of data in a quantitative framework.)

Why model the data at all? What not use the data as the model? The data can fit the data perfectly. In a predictive regression context, if the predictors for a new case match perfectly with the predictors for a case in the data, then the dependent value for the case in the data set can be used as the prediction. The generalization of this uses some function of nearest neighbor values to determine the prediction. In some cases nearest neighbor models can be hard to beat.

There are problems with and limitation to the data as the model approach. There many be many quality problems with the data. The data can be too spotty so that new cases are often too far observed the training set data. Important variables for matching cases can be missing and the best variables for matching many not have been selected for use.

Data can soon become dated for variety reasons. The most obvious reason is that the most interest is often about in the most current data. For example in the phone call fraud detection context people will most likely be gone unless identified using recent data. Another reason for data becoming dated is evolving technology. NASA has new satellites sensors producing better data with higher spectral resolution. If there is too much data to analyze why bother with the inferior data?

4. Cross Validation and Model Fitting

Leave-one-case-out diagnostics can help to identify influential cases. The class regression diagnostics assignment includes a diagnostic that indicates how much the predicted value changes when each case in removed from the model.

The more general idea is to set aside cases for validation purposes. Then the validation cases predicted values using the model based on the rest of the data can be compared against their observed values. In 10 fold cross validation the data is stratified into 10 sets of roughly equal size. Applying the cross validation procedure to each set in turn yields observed and predicted values for all the cases. The discrepancy between observed and predicted values can be used to better characterize the prediction error when applying the model to new data with predictors. The complete data model residuals are likely to capture less of the uncertainty in jumping to a new data sets with its own collection of idiosyncrasies.

Modeling fitting can be based on reducing the cross validation error. This can help to avoid over fitting in the in the 10 models.