Lecture Notes in Quantitative Biology

Exploratory Data Analysis -- Introduction

Chapter 19.1

Revised 25 November 1997
Revised June 2024

ReCap: Correlation
Exploratory Data Analysis
- What is it ?
- Exploratory vs confirmatory analysis
- EDA and inference
- Characteristics of EDA
- 4 Tactics
- Execution
- Box and arrow diagrams
- Examples: Dropping variables.
  - seabirds and El Niño
  - phytoplankton on Georges Bank

on chalk board

Today: Exploratory data analysis. Introduction.

Wrap-up

Purpose of exploratory analysis is to discover pattern.
4 tactics: drop, add, combine variables, discover variables via residuals
Today, we looked at distinction between exploratory and confirmatory.
We also learned about box and arrow diagrams, to simplify and bring out the relation of variables.

EDA--What is it ?

The purpose of exploratory data analysis is to discover or uncover pattern.

We used exploratory data analysis to find a better model. Based on repeated rescaling of the ltot and lthor, repeated plotting, until a linear looking relation was found.

Exploratory data analysis is iterative in nature. There are successive rounds of analysis to discover the pattern.

Exploratory vs Confirmatory Analysis

Equations are used in exploratory analysis of data
What is the best model, given the data ?

Equations are used in confirmatory analysis of data
What can we conclude about the data, given a model ?

Confirmatory and exploratory analyses differ in a number of ways.

Confirmatory Analysis	Exploratory Analysis
What decision can be made ?	What is the appropriate model ?
How certain can we be ?	What is data telling us ?
What are values of parameters ?	What is structure of model ?
Sample	Batch of data
ONE use of a sample (data-grinding, otherwise)	Repeated use of a batch
Single analysis	Iterative search for pattern
p-value = ?	Explained variance = ?
Yes/no decision	Best model
Residuals acceptable ?	Residuals show pattern ?
Experimental design	Factor analysis

This material on board before class, off to side, to save time writing while explaining it. Keep on board for reference later in lecture, bringing out distinction between exploratory (iterative) vs confirmatory (one analysis).

Some statistical methods are more appropriate in exploratory than in confirmatory settings. For example, factor analysis and similar types of multivariate analyses are more suitable for exploratory then confirmatory analysis.

Abuses of statistical analysis are often related to use of confirmatory methods in an exploratory analysis. For example, repeated use of data set is legitimate in exploratory analysis, not legitimated in confirmatory analysis, where it becomes data grinding.

In any statistical analysis, whether confirmatory or exploratory, it is important to remember that pattern does not equal cause. A pattern, no matter how statistically significant, is only a clue to cause. Correlation, no matter how significant statistically, does not prove causation. A "treatment" effect in an experiment may be significant, but not due at all to the treatment administered.

EDA and inference.

1. Exploratory data analysis uses many of the same procedures as those in statistical inference. But it is important to recognize that there are important differences between EDA and statistical inference.

2. In EDA we are asking: What is the model ? (or pattern).

In confirmatory analysis we are asking: Given this model, what can we conclude at the α certainty level ?

3. Because of this difference in focus, the nature of inference is different.

In confirmatory analysis we are using a very narrow form of inference. We are inferring the true value of a parameter (mean, slope, etc) that relates one quantity to another. We are inferring this true value from a sample. The sample should be representative if we want to say anything about the population.

In EDA we are using a much broader form of inference. We are inferring the form of the relation between two quantities, rather than the specifics of a given relation. We are making this inference based on a batch of data. We are not basing our search for pattern on a formal relation between sample and population. We are probably assuming that the batch is "typical" so that the pattern we discover is worth stating as an hypothesis about other such batches. The relation between the batch and the population in EDA is much looser than the relation between sample and population in confirmatory analyses. With EDA we are not trying to confirm pattern, based on a sample. Instead, we are trying to discover a pattern that is worth testing in a confirmatory analysis.

4. The distinction between confirmatory and exploratory analysis has been emphasized here because too often the two are confused, leading to poor practice in the use of confirmatory methods. The use of confirmatory analysis in an exploratory setting usually results in data grinding: repeated analysis of the same batch of data. The statistical decsisions that arise from repeated analysis are meaningless--the error rate is not 5%, even though this might be the stated α level.

5. Of course to some degree we have to use some of the same elements of EDA in confirmatory analysis. We saw this in the generic recipe for hypothesis testing

with the GLM. We examined the plot of residuals versus fits to make sure that our model was not wildly off the mark and completely inappropriate. If the model is unsuitable, we go back and find a better model. But the focus was on declaring the decision, not on finding the model. In EDA the focus changes from declaring a decision to finding the pattern.
6. So exploratory approaches are used in model revision leading to confirmatory analysis. However, we have avoided declaring a sequence of statistical decisions. Only one decision is declared, even though we may have had to revise our model along the way. If we use statistical criteria to revise model, then confirmatory analysis is no longer valid. The p-values will not be correct.

Characteristics of EDA

The goal of EDA is discovering the model rather than making a decision, given the model.

EDA relies strongly on graphical analysis, because of the tremendous capacity of the eye to pick out pattern. Statistical summaries easily miss pattern that is evident in a plot.

EDA uses a batch of data (which has an unknown relation to population). It does not use a sample with known relation to a larger population. There is no formal statement of the population from which a sample is drawn.

EDA is iterative in nature. Analysis continues until a satisfactory description of pattern has been obtained. Satisfactory might mean that no more pattern is evident in the residuals. Satisfactory might mean the simplest model (fewest terms), rather than the best fitting model.

The goal of EDA is to discover pattern, to separate pattern from noise. So the analysis of residuals is an important part of EDA. Residuals can be examined, for clues to pattern. This is what we do when checking the residuals for bowls or arches.

Draw data points that would normally be fitted by a straight. Then draw squiggly line that passes through every point, rather than a straight line.

We are also interested in simplification. Obviously we can draw a complex line that passes through every point. But we are more interested in a general description of pattern that capture the most important features. We are usually less interested in a model so realistic as to describe every detail, at the expense of the larger picture. This is what we did in the oyster triploidy example (successive approximations).

EDA uses a screening criterion rather than a fixed tolerance for Type I error. This criterion can be called α, but it is not a formal cutoff for declaring a decision, which is the meaning of α in a confirmatory setting. The screening criterion α can be set at 10%, 15%, even 20%, in order to reduce Type II error, the chance of missing a pattern. Type II error rises if the screening criterion is set at a small value (e.g. less than 5%).

In EDA the screening criterion is used for the sake of consistency and ease of analysis (what should I call a pattern?). It is not used in relation to an estimate of Type I statistical error.

EDA -- 4 Tactics

The strategy is to find pattern. The tactics are many.

Because we are not worried about correct estimates of Type I error, there are fewer constraints on what is correct or not correct. We can use any procedure we like, provided we do not claim we have estimated Type I error.

This opens up many different ways of proceeding.

Can start with complicated model and drop variables. Start with all measured variables, and move toward simplification by dropping variables.

Can start with simple model and add variables. Move from simple to more elaborate description of pattern.

Can discover variables by examining residuals. An example is Tukey's onion peeling technique. (cf J. Tukey *Exploratory Data Analysis*.

1977. Addison Wesley.). A pattern is stated in the form of a model, the residuals are computed and examined, then pattern in the residuals is used to develop a revised model. This process of peeling away layers of pattern continues until the residuals show no pattern.

Can combine variables in new ways. Start with all measured variables, and more toward sipmlification by creating new variables that are combinations of the first set of variables.

Multivariate methods are examples of this.

Some people like to proceed by adding variables one at a time, perhaps using Tukey's onion peeling technique. Others like to start with all the variable quantities, and proceed by dropping variables or erasing arrows in the diagram. And there are a host of techniques for combining variables to arrive a simpler description of pattern. These include the many elaborations of correlation analysis, including multivariate techniques such as factor analysis.

There is an unfortunate tendency to adopt whatever technique is at hand. An example is over reliance on multivariate analysis, a fairly elaborate and powerful technique developed in psychology, and used in some areas of biology (especially ecology) to combine variables as a clue to underlying processes. Perhaps the best guideline is to keep in mind that there are many of ways of proceeding, and not to get locked into the use of one tactic, such as relying exclusively on dropping variables one at a time, or relying only on onion peeling, or only on multivariate methods to combine variables.

EDA -- Execution

EDA relies heavily on graphical methods. The eye is tremendously good at picking out pattern. Typically statistical summaries are used in conjunction with graphical methods.

Some of the elements of good quantitative practice in the GLM recipe continue to apply.

Define all quantities that are used
Procedure statement
Name and Symbol
Values with Units

Identify response and explanatory variables.

Decide whether to undertake exploratory or confirmatory analysis, stating reasons for choice.

State screening criterion, rather than criterion for significance, to distinguish EDA from confirmatory analysis.

Separation of response and explanatory variables has been relatively easy up until now because we have been using text examples. In data situations where exploratory analysis is appropriate, this will likely be more difficult.

One source of difficulty is that a single variable may be both an explanatory and a response variable. For example, ocean warming due to an El Nino event may alter the abundance of prey, which may in turn alter breeding success of seabirds. Fish abundance is both an explanatory and response variable.

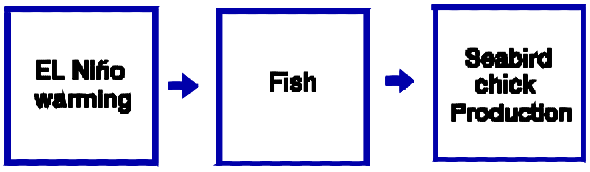
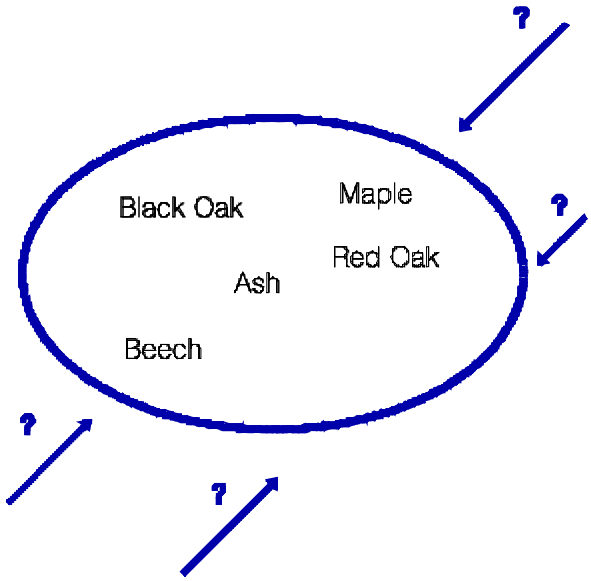


Fig L26f
3 boxes: El Niño warming-->fish-->birds

Another source of difficulty is that we may have a large number of variables with little information on causal ordering. For example, we may be interested in the co-occurence of tree species in relation to environmental factors, but only have information on the abundance of several species at a number of locations. In this case we have as many response variables as tree species. We have no measured explanatory variables. We wish to use pattern of co-association of the response variables as clue to explanatory variables.



EDA -- Box and Arrow Diagrams

Box and arrow diagrams are a useful device in exploratory data analysis. These show how quantities might be related. There are several possible relations among quantities in a box and arrow diagram.

The convention used is to draw boxes around each variable, then draw an arrow from explanatory variables to response variables.

arrow connecting wind stress τ m^{-2} s^{-2}
to upwelling v upward = m day^{-1}
to primary production Pprod = gC m^{-2} day^{-1}

Draw example, wind stress from wind speed and direction

One important relation is that of a quantity calculated from another quantity.

Another important relation is known causal connection. Is there any evidence from other studies that one variable is causally related to another ? Or is there evidence that two quantities are related via a common causal variable (which has perhaps not been measured). If the latter is the case, the common causal variable can be included in the diagram, even though it cannot be analyzed for lack of measurement.

It also helps to separate known from suspected causal relation. A simple device is to put a question mark on an arrow that represents a suspected causal relation. Two question marks can be used to distinguish potential connections from suspected connections for which there is poor evidence.

Another useful device is to place several variables in a single box.

wind stress, light, nutrients in one box,
grazing pressure in another box,
connected to primary production

Example. Numbers of trees in quadrats, 6 species, 20 quadrats.

Show data matrix. List variables on board, boxes around the whole lot.

Arrows to this box (multiple unknown factors).

Example. Gordon Riley interested in aquatic productivity of Georges Bank.

Measurements of microbial (phytoplankton) concentration.
zooplankton concentration
light
nutrients (nitrates, phophates)

List variables on board, as boxes
Ask for arrows. Sequence likely to be:
light --->phyto
nutrients (P N) ---> phyto
phyto---> zoopl
Then add some arrows based on research findings
zoopl ---> nitrate (excretion)
phyto ---> light (shelf shading)
(50 years of research history in 50 seconds)
Compare number of arrows drawn to potential number.
5 * 4 / 2 connections * 2 arrows/connections = 20 arrows
Hence simplified to 6 arrows, from 20 possible.