Lecture Notes in Quantitative Biology

Exploratory Data Analysis -- Adding Variables

Chapter 19.4 (added in 1994)

on chalk board

ReCap
EDA - Adding variables
Example.

In 1996, took 15 minutes because diagram from previous lecture was still on the board.
In 1997, done in 5 minutes (too quick), in same lecture as drop variable example.

ReCap. EDA. What is it ? It is a combination of graphical and formal analysis with the goal of discovering the "best" model.

EDA and inference. The inference from the model to a larger population is much looser than in formal statistical inference. The model is the "best" description for a batch of data. It may possibly apply to some larger population from which that batch was drawn. The result of the exploratory analysis represents a best first guess. It can be used for subsequently hypothesis testing.

Characteristics of EDA.
- Purpose is to discover pattern, so EDA often uses analysis of residuals.
- Uses a batch of data, rather than sample in the strict sense.
- Iterative.
- Uses a screening criterion rather than a significance level α

Tactics Combine variables (multivariate analysis, extension of correlation)
- Drop variables (example yesterday)
- Add variables (example today)

Execution. Elements of good quantitative analysis still 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, state reasons, use screening criterion or significance level, as appropriate.
- Box and arrow diagrams often useful.

Example of seabirds and El Niño to illustrate simplification of model by dropping variables.

Today. Further analysis of seabirds in relation to El Niño, to illustrate tactic of adding variables.

1. Define quantities.

Add diagrams to lecture notes

Same as Ch19.3.
RSPECO = Reproductive success in Pelagic Cormorant
RFdiet = proportion of juvenile rockfish in diet fed to chicks by Common Murre. This is used for other avian species because it was found to be closely related to a measure of rockfish abundance in regular trawl surveys in the region.

AL = index measuring the warming effect on surface waters by low pressure systems centred to the north, in the Aleutian Islands.
ALs = spring-summer cooling phase
ALw = winter warming phase

SO = index measuring the warming effect on subsurface water by oscillating air pressure effects (southern oscillation) in the tropical Pacific.
SOs = spring summer cooling phase
SOw = winter warming phase

2. Box and arrow diagram

Arrows from 2 boxes sAL and wSO to 3 boxes: RSPECO, UW and SLfeb. Add arrow from UW to RSPECO, but not from SLfeb to RSPECO (no direct effect).

These connections thought to be due to effects of upwelling on production and availability of an important prey item, juvenile rockfish.
Should prey be added ?

Does adding this variable increase the level of explained variance?

3. Rationale for exploratory analysis.

Still looking at the question of what is the best model of seabird reproductive success. So continue using exploratory analysis.

4. Criterion and Procedure

The most effective way to evaluate whether juvenile rockfish should be added to model is to take the residuals from the model based on ocean climate variables, then plot these residuals against measure of rockfish. Simply looking at this plot will reveal a lot. It will provide clues as to the shape of the relation, if any between seabird reproductive success and supply of juvenile rockfish, corrected for the effects of oceanic climate on food supply.

The authors of the article did not do this, nor do they supply the data in an appendix, which would allow us to re-analyze the data.

The authors continued to use same procedure and criterion: linear regression and level of explained variance R^2 to evaluate whether to add food to the model.

5. Model

RFdiet added to determine the degree to which physical variables could be surrogates of availability of this prey item. If there is no improvement from adding this, then physical variables can act as surrogate measures of reproductive success. The authors of the article chose to work again by simplifying the full model.

$$\text{Log}(1 + \text{RSPECO}) = \beta_o + \beta_{sSO}\text{sSO} + \beta_{SOw}\text{SOw} + \beta_{ALs}\text{ALs}$$

$$+ \beta_{ALw}\text{ALw} + \beta_{UW}\text{UW} + \beta_{SLfeb}\text{SLfeb} + \beta_{RFdiet}\,\text{RFdiet} + \text{res}$$

There are now 8 possible factors.
The number of possible models has grown to 255.

k	formula	number of models	
1			8
2	8!/2!6!	=	28
3	8!/3!5!	=	56
4	8!/4!4!	=	70
5	8!/5!3!	=	56
6	8!/6!2!	=	28
7	8!/7!1!	=	8
8			1
		total:	255

6. Revised model

Same criterion: examine all models for highest R^2

$$\text{Log}(1 + \text{RSPECO}) = 0.047 - 0.172\text{SOw} + 0.301\text{ALw} - 0.002\text{SLfeb} + 0.008\ \text{RFdiet}$$

$R^2 = 56\%$

Compare new model to the previous model (ocean climate variables only).

$$\log(1+\text{RSPECO}) = 0.577 - 0.172\text{SOw} - 0.494\text{ALs} - 0.012\text{UW} - 0.003\text{SLfeb}$$

$R^2 = 51\%$

The new model is similar to the previous, with slight changes. It has two large scale variables; one is the same as before. The other is different: the winter phase of the Aleution Low (ALw) instead of the summer phase (ALs). The new model has one local index of upwelling (SLfeb), rather than two. The addition of rockfish increased the explained variance R^2, but not substantially more than first model. Hence the first model (only physical variables) explains reproductive success nearly as well. The lack of substantial improvement indicates that prey other than rockfish, and the effects of physical variables on these prey are important in this species, the Pelagic Cormorant.

7. Revise diagram

The model based on ocean variables works as well as the model based on ocean varables + juvenile rockfish. This indicates that juvenile are not the key prey item.

Box labelled "food" rather than "RFdiet" added to first diagram.

8. Carry out with other species, compare results.

Substantial improvement observed in only one species, the Pigeon Guillemot. $R^2 = 47\%$ for model with physical variables. $R^2 = 92\%$ when diet included. This suggests that availability of one prey species, rockfish, is of primary importance, and that ocean climate effects on prey of this species of seabird are of secondary importance.

Add RF_{diet} to diagram.

Substantial improvement was not observed in the other 5 seabird species, including the detailed example above, Pelagic Cormorant. For these 5 species, the availability of one prey species, rockfish, is less important than ocean climate, presumably through effects on the availability of all prey, rather than just one species. This is consistent with many (though not all) studies of seabird reproductive success in relation to diet, which show that seabirds can utilize a variety of prey to feed and successfully fledge chicks. This exploratory analysis suggests that ocean climate alters the availability of the entire suite of prey species, which in turn alters reproductive success.