Exploratory Data Analysis. Dropping Variables

Chapter 19.3 Revised 26 November 1997

ReCap: EDA
4 Tactics
Examples: Dropping variables.
seabirds and El Niño
phytoplankton on Georges Bank

on chalk board

Recap (from 19.2)
Purpose of exploratory analysis is to discover pattern.
4 tactics: drop, add, combine variables, discover variables via residuals
We looked at distinction between exploratory and confirmatory.
We used box and arrow diagrams,
to simplify and bring out the relation of variables.

Today: Exploratory data analysis. Dropping and Adding Variables.

Wrap-up Purpose of exploratory analysis is to discover pattern.
4 tactics: drop, add, combine variables, discover variables via residuals
Today, we looked at one of these, dropping variables.

EDA -- Example with dropping variable, using stepwise multiple regression

Written in 1994.

Data. Reproductive success of 6 species of marine bird. Abundance of one species of prey. 7 measures related to oceanic conditions that affect prey abundance in the California Current system. Data collected over 21 years.

From: Ainley, D.G., J. Norton, W.J. Sydeman. 1995. Apex predators indicate interannual negative and positive anomalies in the California Current food web. *Marine Ecology Progress Series* 118: 69-79

These authors investigated the relation of seabird reproductive success to ocean climate effects. They used multiple regression. The results are better considered exploratory than confirmatory.

1. **Define quantities**, with symbols and names. (procedural statements in the paper).

RSBRCO = Reproductive success in Brandt's Cormorant
RSPECO = Reproductive success in Pelagic Cormorant
RSWEGU = Reproductive success in Western Gull
RSCOMU = Reproductive success in Common Murre
RSPIGU = Reproductive success in Pigeon Guillemot
RSCAAU = Reproductive success in Cassin's Auklet

Reproductive success in colony on Farallon Islands (off coast near San Francisco) measured as chicks fledged per pair that attempted breeding).

RFdiet = proportion of juvenile rockfish in diet fed to chicks by Common Murre.
RFtrwl = abundance of juvenile rockfish in standardized trawls in this area.

AL = index measuring the warming effect on surface waters by low pressure systems centred to the north, in the Aleutian Islands.

SO = index measuring the warming effect on subsurface water by oscillating air pressure effects (southern oscillation) in the tropical Pacific.

These two indices measure atmosphere-ocean dynamics that influence the production and availability of seabird prey, including juvenile rockfish.

1. Define quantities (continued)

These two indices were separated into seasonal components: a spring-summer cooling phase (sAL and sSO) and a fall-winter warming phase (wAL and wSO)
The Aleution Low (AL) and Southern Oscillation (SO) are global scale influences on upwelling in the California Current. Three additional measures of upwelling intensity in the California Current were also used in the exploratory analysis. These were

UW = an index of upwelling intensity based on wind stress along the coast in January and Febuary.

Tmar = sea surface temperature in March (drop depends on upwelling)

SLfeb = sea level at the coast in February (drop depends on upwelling).

2. Identify response and explanatory variables.

Here is box and arrow diagram.

Fig L26h. Arrows from 4 large scale variables to box with three smaller scale upwelling variables. Arrow from this box to prey box, RFtrawl occupying one of the compartments of this box. Arrow from this box to diet box, with on compartment labelled RFdiet. Arrow from this box to each of the 6 bird species.

3. Rationale for exploratory analysis

The strength of causal connections in this diagram is likely to vary considerably. The strengths are unknown, so we will use exploratory analysis to find the best combination of variables as clue to underlying process.

4. Procedure and Criterion. The authors used maximum explained variance, rather than using a set p-value. Explained variance is a measure of the overall model, allowing comparison among models for screening.

The authors used stepwise multiple regression to partition variance into an explained (variance due to the model) and unexplained portion (variance of the residuals). The ratio of the sum of squares SS_{model} / SS_{tot} is called R^2, the explained variance. The procedure is to compute the explained variance for all possible models, then choose model with best (largest) R^2.

4. Procedure and Criterion (continued)

Here is a tally of all possible models, for 7 explanatory variables. k is the number of factors in the model.

$4! = 4 \cdot 3 \cdot 2 \cdot 1$

k	formula		number of models
1			7
2	7!/2!5!	=	24
3	7!/3!4!	=	35
4	7!/4!3!	=	35
5	7!/5!2!	=	24
6	7!/6!1!	=	7
7			1
		total:	133

The model with the maximum R^2 is then chosen from these 133 possible models. This will not necessarily be the full factor model because R^2 is adjusted for the presence of other factors in the model.

5. Formal model

$$\text{RSPECO} = \beta_o + \beta_{sSO} . \text{sSO} + \beta_{wSO} \cdot \text{wSO} + \beta_{sAL} \cdot \text{sAL} + \beta_{wAL} \cdot \text{wAL}$$
$$+ \beta_{UW} \cdot \text{UW} + \beta_{Tmar} \cdot \text{Tmar} + \beta_{SLfeb} \cdot \text{SLfeb} + \text{res}$$

6. Revised model as reported by Ainley *et al.*

$$\log(1+\text{RSPECO}) = 0.577 + -0.172\ \text{wSO} + -0.494\ \text{sAL} +$$
$$-0.012\ \text{UW} \quad -0.003\ \text{SLfeb}$$
$$R^2 = 51\%$$

The model contains two large scale forcing variables, the winter phase of the Southern Oscillation (wSO) and the summer phase of the Aleutian Low (sAL). It also contains two local indices of upwelling, wind stress (UW) and sea level (SLfeb). [Sea level drops several cm during upwelling, as the water is pulled away from the coast].

Critique.

1. Did the model capture the relation of RSPECO to explanatory?
   No evaluation (bowl / arch criterion on residual vs fit plot.
2. The result depends on the selection criterion:
   best fit (this case) versus some other criterion
3. Log transform used because recommended to ‘meet assumptions’ for count data.
   Assumption not checked.
   Log transform requires adding an arbitrary constant (+1 in this case)
   where there are zero values.
4. 21st century statistical packages allow us to estimate exponential relation without having to take log transform.

$$\text{RSPECO} = e^{\mu} + \in$$

where $\mu = \beta_o + \beta_{sSO} . \text{sSO} + \beta_{wSO} \cdot \text{wSO} + \beta_{sAL} \cdot \text{sAL} + \beta_{wAL} \cdot \text{wAL} + \beta_{UW} \cdot \text{UW}$
$+ \beta_{Tmar} \cdot \text{Tmar} + \beta_{SLfeb} \cdot \text{SLfeb}$

The results reported depend on the choices made. Better considered exploratory.

7 Revised diagram

Here is a tabular summary of the variables that were retained.

	RSPECO
sSO	
wSO	x
sAL	x
wAL	
UW	x
Tmar	
SLfeb	x

Fig L26I. All 7 factors on left, PECO box on right. Connect factors to PECO box with 4 arrows. Then add box in middle of each arrow, labelled 'prey.'

The box and arrow diagram is then redrawn in simpler form, based on the results of this exploratory analysis.

8. Compare to other species. Different results were obtained for the other 6 seabird species.

	RSBRCO	RSPECO	RSCOMU	RSPIGU	RSCAAU
sSO			x		
wSO	x	x	x	x	x
sAL	x	x			x
wAL	x		x		x
UW		x	x	x	x
Tmar	x		x	x	x
SLfeb		x	x	x	

Note that wSO shows up in all species.

EDA -- Example with dropping variable, using multiple regression

From 22 November 1993.

Data. Phytoplankton abundance in net tows from Georges Bank, in the 1950s. Gordon Riley used multiple regression to identify which variables were of primary importance. This example shows how Riley proceeded.

Data: Concomitant measurements on 5 variables at a series of locations on Georges Bank

Light,
Nutrient concentrations (Nitrogen, Phophorus)
Zooplankton density (grazers)

1. Define quantities, with symbols and names.

Phy = chlorophyll units
L = secchi disk depth (greater the depth, the more the light)
N = nitrogen per ml per liter of water
P = phosphorus per ml of water
Z = zooplankton mass per ml of water.

2. Known or suspected relation of variables. Draw as box and arrow diagram.

L---> Phyto production--->Phyto density

N--->Phyto production--->Phyto density

P--->Phyto production--->Phyto density

Z-->Phyto loss--->Phyto density

This is a fairly simple situation. Four variables suspected to affect one variable.

3. State rationale for exploratory analysis

   Purpose was exploratory: Which effects were important ?
   Many of the explanatory variables covary, and so:
   Is there significant relation of P to each of the other variables, controlling for the remaining variables ?
   Is there significant relation, controlling for colinearity of explanatory variables.

4. Verbal model: phytoplankton growth depends on light, nutrients, zooplanktonic grazers

5. Graphical display.

   Phy against L, Phy against N, Phy against P, Phy against L.

   2-d graphs for explanatory variables:
   P against N,
   P against L,
   P against Z,
   N against L,
   N against Z,
   L against Z.

   Turn these graphs 45° clockwise, to erase the idea of causal ordering, of X--->Y

   Then 3-d graphs.
   Phy (vertical) against N and P,
   Phy (vertical) against N and L
   Phy (vertical) agaisnt N and Z
   Phy (vertical) against P and L
   Phy (vertical) against P and Z
   Phy (vertical) against L and Z

   That was 3-d display. How about 4-d display ? Can it be done?
   Phy against N, P, and L
   Phy against N, P, and Z
   Phy against P, L, and Z
   Phy against N, L, and Z

   Yes, imagine plot of Phy against N, P (3-d) that changes on computer screen as Light goes from low to high.
   Similarly, imagine plot of Phy against N, P (3-d) that changes on screen as Zooplankton goes from low to high.

   How about 5-d display ?

   It turns out this can be done

   Imagine a whole row of computer screens, each displaying
   Phy against N and P, changing as light goes from low to high
   screens on left, as you look at them, show low zoopl levels
   screens in middle show intermediate Z
   screens on right show high Z

6. Calculate pattern. Use linear model in this case, because each effect thought to be linear.

Begin with over all test, then see if some terms can be dropped.

$$\text{Phyto} = \beta_0 + \beta_L \cdot L + \beta_N \cdot N + \beta_P \cdot P + \beta_Z \cdot Z + \epsilon$$

order does not matter because using an overall test to start.

Using GLM so table partitioning of variances

Source	df	SS	MS	F
Model	4			
Residual	n−5			

The p-value turned out to be greater than a screening criterion of 5%, so next step is to isolate which terms are more important. Going to try to drop variables because we have reason to think that all variables affect growth, but we suspect that some variables may be far more important than others. Going to use Type III sum of squares. Ie, is there a partial regression, after taking into account the other 3 variables. Partial regression can be represented as a plot of the residuals from a model with the other terms present.
Type I same as Type III SS for only the <u>last</u> term in the model.

$$\text{Phyto} = \beta_0 + \beta_{L.NPZ}\, L + \beta_{N.LPZ}\, N + \beta_{P.LNZ}\, P + \beta_{Z.LNP}\, Z + \epsilon$$

Sum of squares. Partitioning depends on order in model.

Picture of	$\text{Phyto} = \beta_0 + \beta_L L$
Take residuals to obtain picture of	$\text{Phyto} = \beta_0 + \beta_L L + \beta_{N.L} N$
This is not the same as	$\text{Phyto} = \beta_0 + \beta_N N$
Nor the same as	$\text{Phyto} = \beta_0 + \beta_L L + \beta_N N$
	Ie, $\beta_{N.L} \neq \beta_N$

Table (Type 3 SS)

Source	df	SS3	MS	F	p
$\beta_{L.NPZ}$	1				
$\beta_{N.LPZ}$	1				
$\beta_{P.LNZ}$	1				
$\beta_{Z.LNP}$	1	(same as above)			
residual	n - 5				

MS/residual = F

Result was that none can be dropped. The p-value for all terms was above the screening criterion of 5%, so write model with all terms.

Extra, from 22 Nov 1993

7. Calculate difference between pattern and observation at each step

8. Examine for residuals for pattern.

9. If no pattern in residuals than write simplest model

10. There is a pattern in the residuals try something else