Collinearity in Multi-variate Analyses

- *Objectives:*

Showcase the problem of collinearity

Discuss rationale / use of partial correlation analysis

Discuss rationale / use of partial regression analysis

Illustrate the application of both approaches

Collinearity (or Co-linearity)

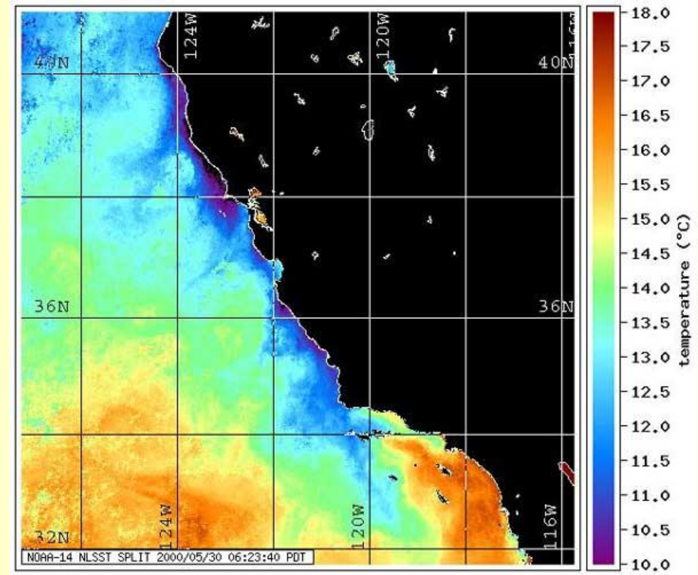
- Collinearity is defined as the existence of a linear relationship between **two** explanatory variables.
- Multi-collinearity refers to the collinearity of **more than two** explanatory variables.
- Implies redundancy in the set of variables. This redundancy can render the numerical methods used to solve regression equations ineffective.
- Implies that significant linear correlations exist between variables, making it difficult (or impossible) to estimate individual regression coefficients reliably.
- Two practical solutions to this problem are:

  to remove some variables from the model…
  or to combine them (using Multivariate Statistics).

Multiple Collinearity

Table 1. Correlation coefficients (r-values) for relationships between oceanographic variables sea-surface temperature (SST), sea-surface salinity (SSS), wind speed (WSP), thermocline depth (TDPT), and thermocline slope (TSLP). Sample size (number of 15 min transects) was 2161

	SST	SSS	WSP	TDPT
SSS	−0.588			
WSP	−0.111	−0.056		
TDPT	−0.208	0.083	0.184	
TSLP	0.589	−0.347	0.051	−0.210



(Oedekoven et al. 2001)

Onshore – Offshore:

- Depth
- Distance to shore
- Productivity

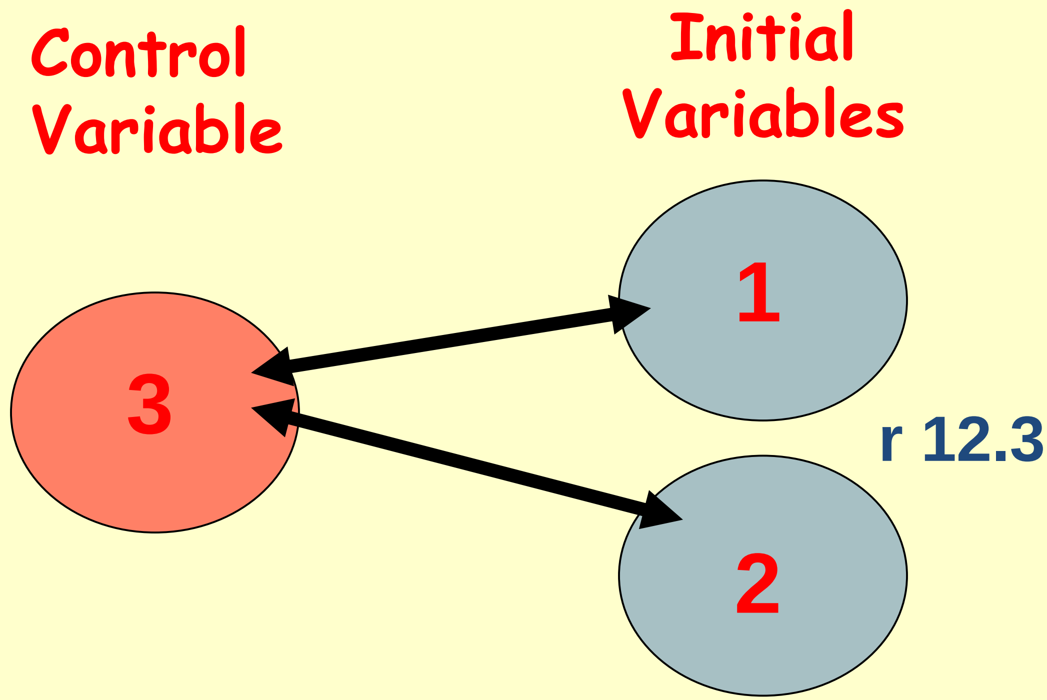
Latitudinal Gradients:

- Water Temperature
- Wind Speed
- Seasonality

Partial Correlation - Application

Zero-order correlation (two initial variables).

A control variable is used to extract the variance it explains from each of the two initial variables, which are then correlated with each other.



The resulting partial (first-order) correlation (r 12.3) is the correlation between the two initial variables which remains once the influence (variance explained by the control variable) has been removed from each of them.

Partial Correlation - Application

- The technique is commonly used in "causal“ modeling, when dealing with a small number of variables (e.g., 3 – 5).
- Compares the partial first-order correlation (r 12.3) with the original zero-order correlation (r 12).
- Sometimes the third variable (3) has no effect on 1 and 2.
- The third variable (3) can have different influences:
  - a common anteceding cause
  - an intervening variable
  - a suppression variable

Partial Correlation – Interpretation

- Compare magnitude of full and partial correlations
  - If the full correlation is meaningful (large) and the partial correlation approaches 0, the inference is that the original correlation is spurious - there is no direct causal link between the two original variables
  - If the full and the partial correlations are meaningful (large) and the similar in magnitude, the inference is that the original correlation is real - there is a direct causal link between the two original variables

Partial Correlation - Interpretation

- Compare magnitude of full and partial correlations

- If there is no difference between the r coefficients, for the full and the partial correlations the inference is that the third variable (k) has no effect on i and j.

$$\boxed{r_{ij} = r_{ij.k}}$$

- If the partial correlation approaches 0, the inference is that the original correlation is spurious - there is no direct causal link between the two original variables (I and j).

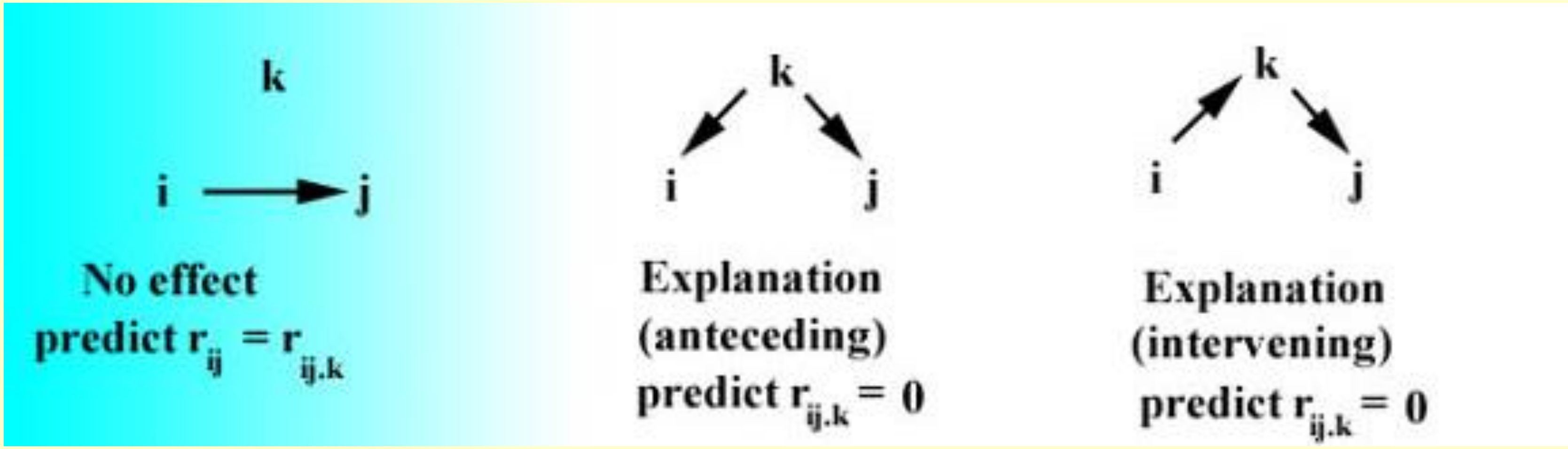
$$\boxed{\begin{array}{c} r_{ij} > 0 \\ r_{ij.k} \sim 0 \end{array}}$$

Partial Correlation – Application

- Compare magnitude of full and partial correlations

K Has No Effect

K Has An Effect

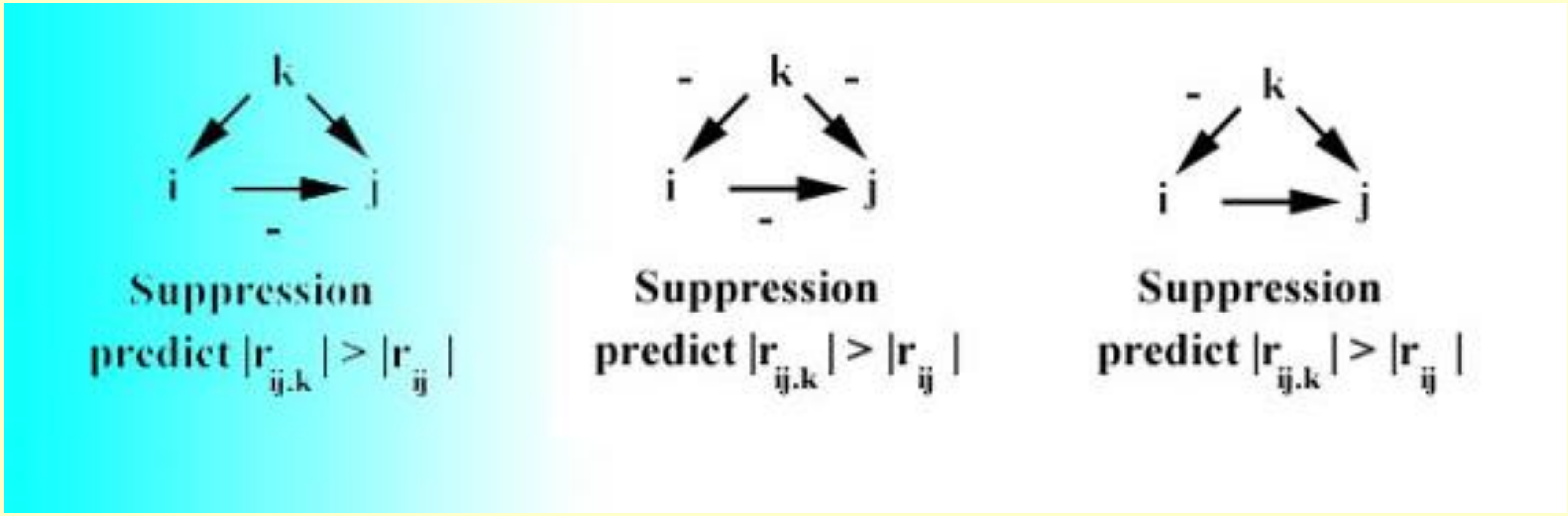


Common Anteceding Cause

Intervening Variable

Partial Correlation – Interpretation

- Compare magnitude of full and partial correlations
  - If the partial correlation < the full correlation, the inference is that the third variable (k) suppresses apparent correlation between two other variables (I, j).



Suppression Effect: Magnitude of the correlation appears weaker, when the third variable is considered

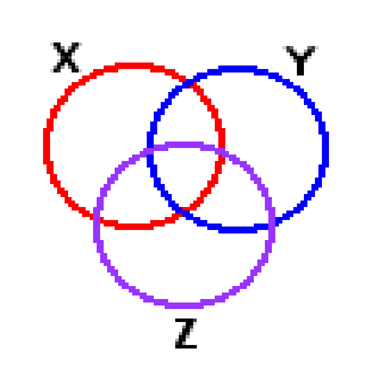
Another Example

- Suppose you measured three variables (X, Y, Z) from N subjects and found the following correlations:

X versus Y: **r** XY = +0.50 **r**^{2} XY = 0.25

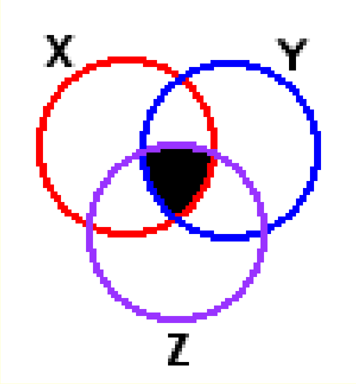
X versus Z: **r** XZ = +0.50 **r**^{2} XZ = 0.25

Y versus Z: **r** YZ = +0.50 **r**^{2} YZ = 0.25



- The value of **r**^{2}, which equals 0.25, implies that for each pair of variables (XY, XZ, YZ) the covariance, or shared variance, is 25% (one quarter of the total).

- The venn diagram illustrates that 25% of the variability in the three variables (X, Y and Z) overlaps (shaded area).



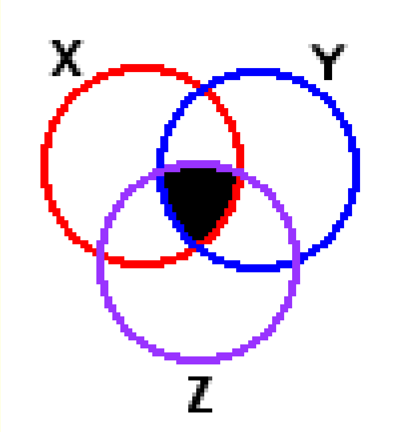
Partial Correlation – 3 Variables

Partial correlation is the correlation of two variables, controlling for influence of one or more other variables.

- Partial correlation allows us to measure the region of three-way overlap and to remove it from the picture.

- This method determines the value of the correlation between any two of the variables (hypothetically) **if** they were not both correlated with the third variable.

- Mechanistically, this method allows us to determine what the correlation between two variables would be (hypothetically) **if** the third variable were held constant.



Partial Correlation - Formulation

- The partial correlation of X and Y, with the effects of Z removed (or held constant), is given by the formula:

$$r_{XY\cdot Z} = \frac{r_{XY} - (r_{XZ})(r_{YZ})}{\mathrm{sqrt}[1 - r^2_{XZ}] \times \mathrm{sqrt}[1 - r^2_{YZ}]}$$

which for the present example would work out as

$$r_{XY\cdot Z} = \frac{.50 - (.50)(.50)}{\mathrm{sqrt}[1 - .25] \times \mathrm{sqrt}[1 - .25]}$$

$$= \boxed{+.33}$$

(Smaller than $r_{xy} = + 0.50$)

Partial Correlation - Formulation

- Partial Correlation of X and Z

$$r_{XZ \cdot Y} = \frac{r_{XZ} - (r_{XY})(r_{YZ})}{\text{sqrt}[1 - r^2_{XY}] \times \text{sqrt}[1 - r^2_{YZ}]}$$

$(r_{xy.z} = + 0.33)$

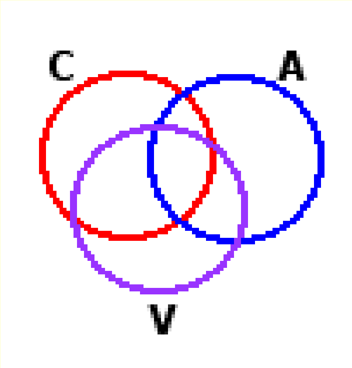
- Partial Correlation of Y and Z

$$r_{YZ \cdot X} = \frac{r_{YZ} - (r_{XY})(r_{XZ})}{\text{sqrt}[1 - r^2_{XY}] \times \text{sqrt}[1 - r^2_{XZ}]}$$

$(r_{xy.z} = + 0.33)$

Partial Correlation – Another Example

➢ Example: Wechsler Adult Intelligence Scale (WAIS) is used to measure "intelligence" beyond the years of childhood. It includes 3 sub-scales labeled C, A, and V.



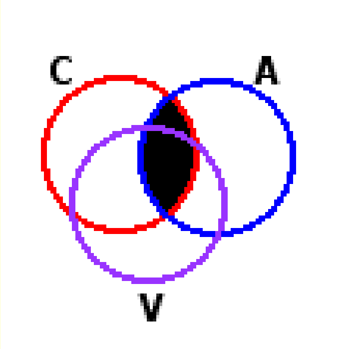
- C: “comprehension”
- A: “arithmetic”
- V: “vocabulary”

The table shows correlations among these 3sub-scales:

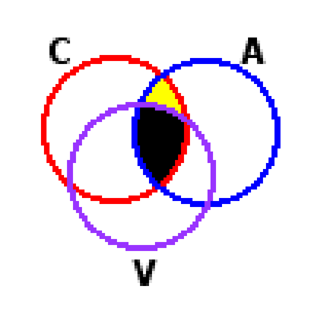
- C versus A: **r**CA = +0.49 **r**^{2}CA = 0.24
- C versus V: **r**CV = +0.73 **r**^{2}CV = 0.53
- A versus V: **r**AV = +0.59 **r**^{2}AV = 0.35

Partial Correlation – Another Example

- While the overlaps are not even, the logic is the same:
- Of the 24% variance overlap in the relationship between comprehension and arithmetic (C & A), a substantial portion reflects the correlations of these variables with vocabulary (V).



- If we remove the effects of V from the relationship between C and A, the partial correlation (r CA.V) will be smaller than the full correlation (r CA).



Remember: Zero-order correlation of C and A: **r**CA = +0.49

Partial Correlation – Another Example

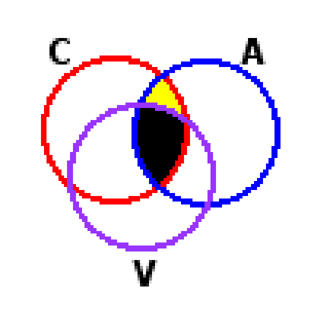
- The correlated overlap is reduced substantially:

$$r_{CA \cdot V} = \frac{r_{CA} - (r_{CV})(r_{AV})}{\textbf{sqrt}[1 - \mathbf{r}^2_{CV}] \times \textbf{sqrt}[1 - \mathbf{r}^2_{AV}]}$$

$$= \frac{.49 - (.73)(.59)}{\textbf{sqrt}[1 - .53] \times \textbf{sqrt}[1 - .35]}$$

$$= +.11$$

Hence $\mathbf{r}^2_{CA \cdot V} = \boxed{.01}$



(Less than r_{CA} = + 0.49)

Partial Correlation – Result

Summary: After removing the effects of V, the correlation between C and A diminishes greatly.

- In most cases a partial correlation of the form $r_{XY \cdot Z}$ is smaller than the original correlation r_{XY}.

- In cases where the partial correlation is larger, the third variable, Z, is termed a **supressor variable**, based on the assumption that it is suppressing the larger correlation that would appear between X and Y if Z were held constant.

Partial Correlation – Example

SOCIAL PARTICIPATION AND VOTING TURNOUT:
A MULTIVARIATE ANALYSIS *

MARVIN E. OLSEN

Indiana University and Uppsala University

American Sociological Review 1972, Vol. 37 (June):317-333 **(Olsen 1972)**

APPROACH: Investigated relationship of voting turnout and participation in church / community activities, voluntary associations and inter-personal interactions.

The social participation hypothesis predicts that:

H1: Participation rate in voluntary associations positively related to voting turnout, whatever nature of the association.

H2: Participation rate in community / church activities also positively related to voting turnout, and remain significant when voluntary association participation is held constant.

Partial Correlation – Example

H3: Participation rate in interpersonal interaction positively related to voting turnout; but
holding constant participation in voluntary associations / churches / community affairs eliminates these correlations.

RESULTS:

Partial correlations for each of these measures, controlling the others, remain significant, indicating that each form of community participation has independent effects on voting.

However, this is not true with informal inter-personal interactions among friends and neighbors.

The mean multiple r with all predictor variables is 0.58.

(Olsen 1972)

Partial Correlation – Example

Cancer incidences in Europe related to mortalities, and ethnohistoric, genetic, and geographic distances

Robert R. Sokal*†, Neal L. Oden‡, Michael S. Rosenberg*, and Barbara A. Thomson*

*Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245; and ‡EMMES Corporation, 11325 Seven Locks Road, Suite 214, Potomac, MD 20854

- Rationale: Differences in cancer rates between populations are affected partly by geographic distances and by ethnic differences. Epidemiologists ascribe these differences to genetic and environmental factors.

- Analyses: Spatial autocorrelation of cancer incidence and mortality rates, ethnohistory, and genetics showed that these four variables are strongly spatially autocorrelated.

(Sokal et al. 2000)

Partial Correlation – Example

To learn more of the interaction of the variables, constructed a path diagram. The three putative causes are labeled ETH, GEN, and GEO.

Path Diagram: Map of associations between all variable pairs:

0 order: Ordinary pairwise correlations

GEO, GEN and ETH are cross-correlated

GEO, GEN and ETH drive other variables

Why is this a one-headed arrow ?

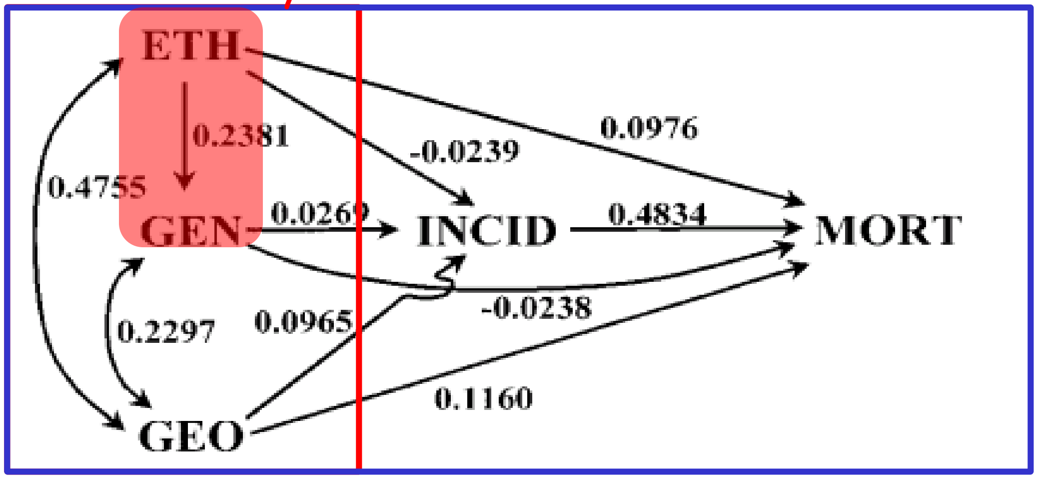


Fig. 2. Path diagram for zero-order partial correlation coefficients (ordinary pairwise correlations) between distances or differences of the indicated variables. Double-headed arrows indicate correlations; single-headed arrows are paths from the base to the tip of the arrows. The numerical values alongside the arrows indicate magnitudes of the correlations or path coefficients.

(Sokal et al. 2000)

Partial Correlation – Example

Path Analysis.

Both ETH-GEO and GEN-GEO are connected by double-headed arrows to indicate remote correlations that we cannot decompose further.

The correlation r(ETH,GEN) is shown as a single-headed arrow ETH -> GEN, because ethnohistoric similarity will lead to genetic similarity, whereas the converse will not hold generally.

Ethnohistoric distances imply not only genetic distances (affected through the path ETH -> GEN), but also cultural differences that directly affect the differences in cancer rates. These are shown as separate single-headed arrows ETH -> MORT and ETH -> INCID.

Partial Correlation – Example

Created four distance matrices:

cancer incidence (INCID) ethnohistory (ETH)

genetics (GEN) geography (GEO)

Computed all zero-order matrix correlations, as well as partial matrix correlations between INCID and ETH and GEN.

- 0 order (INCID, ETH) (INCID, GEN)

- 1st Order (INCID, ETH.GEO) (INCID, GEN.GEO)

- 2nd Order (INCID, ETH.GEN, GEO) (INCID, GEN.ETH, GEO)

(Sokal et al. 2000)

Partial Correlation – Example

- 0 order: GEN seems more important (mean r, n)
- 1st Order: ETH correlations explained by GEO
- 2nd Order: GEN retains explanation power, with ETH, GEO

Table 1. A summary of zero-, first-, and second-order partial correlations of cancer incidence distances (INCID) with ethnohistoric (ETH) and genetic (GEN) distances in Europe

	Males, 45 cancers			Females, 47 cancers		
	ETH	GEN	*n* (ETH > GEN)	ETH	GEN	*n* (ETH > GEN)
Zero order	0.0370*	$\boxed{0.0492^*}$	20	0.0197†	$\boxed{0.0377^*}$	17
First order	−0.0094	$\boxed{0.0470^*}$	16	−0.0264	$\boxed{0.0372^*}$	16
Second order	−0.0057	$\boxed{0.0312^*}$	10	−0.0046	$\boxed{0.0282^*}$	12

Values in columns one, two, four, and five are averages of partial matrix correlation coefficients (11, 19) as follows: zero order in the ETH columns stands for *r*(INCID,ETH), first order is *r*(INCID,ETH.GEO), and second order is *r*(INCID,ETH.GEN,GEO), where GEO stands for geographic distances. For the GEN columns, interchange GEN with ETH. The significance indicators next to the averages are based on Fisher's method of combining probabilities (22). *, $P \ll 0.001$; †, $P = 0.030$. In columns three and six, headed *n* (ETH > GEN), we furnish counts of the number of cancers for which correlation of cancer incidence distances with ethnohistoric distances exceeds that with genetic distances.

(Sokal et al. 2000)

Linear Regression

How to assess correlation between two variables without the effect of another exogenous driver variable (third-variable) ?

Y= a + bX + e   Equation of a Straight Line

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

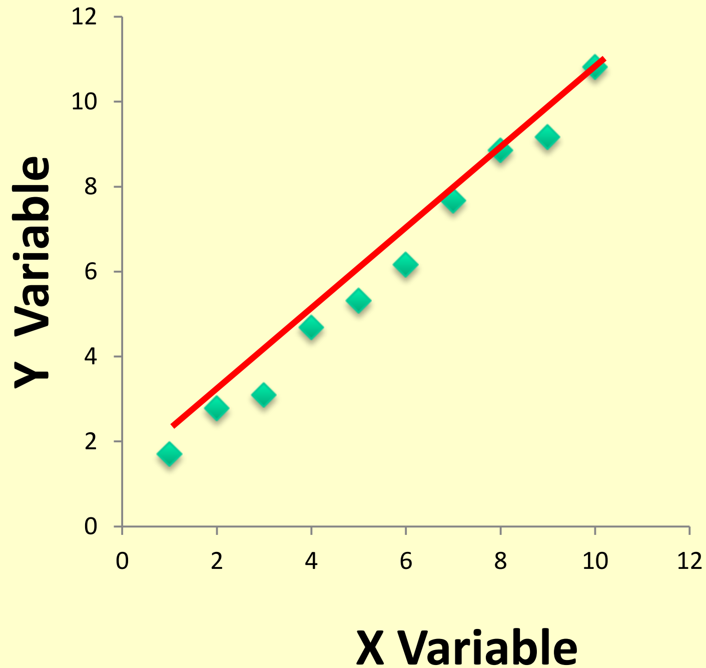
$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

Estimates y values from x values

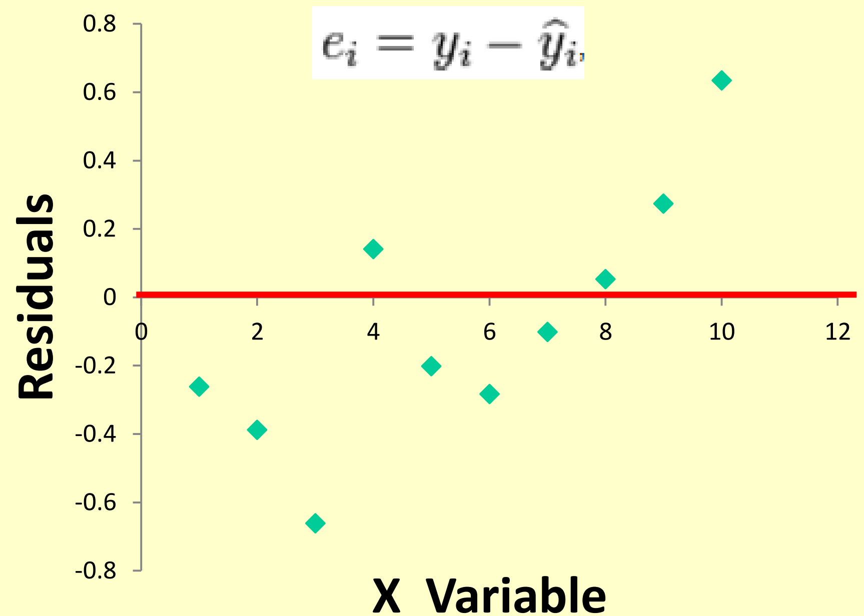
$$\widehat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Linear Regression

Observed X-Y Data Pairs



Y on X Pair Residual Plots



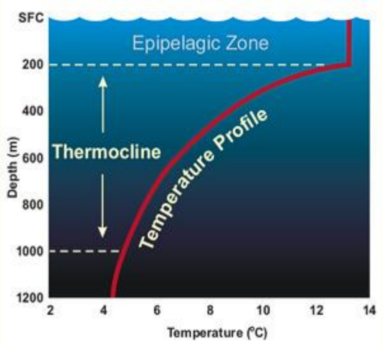
Fit = Sum of Squares Error =
sum of squared deviations
(observed data – model prediction)

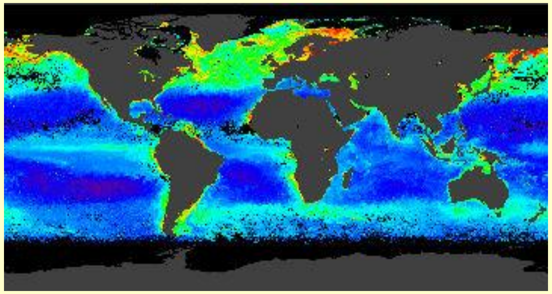
$$SSE = \sum_{i=1}^{N} e_i^2 .$$

Linear Regression

For instance:

Is the number of breeding seabirds on an island more strongly related to thermocline depth OR chl concentration ?





CHL-a Concentration

THERM Depth

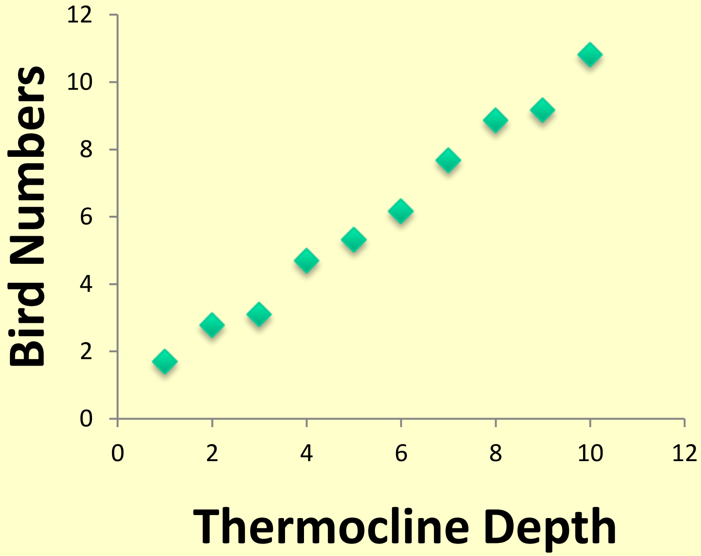
SEABIRD Population

Need to “remove” the effect of THERM on CHL and BIRDS to assess the correlation between CHL and BIRDS?

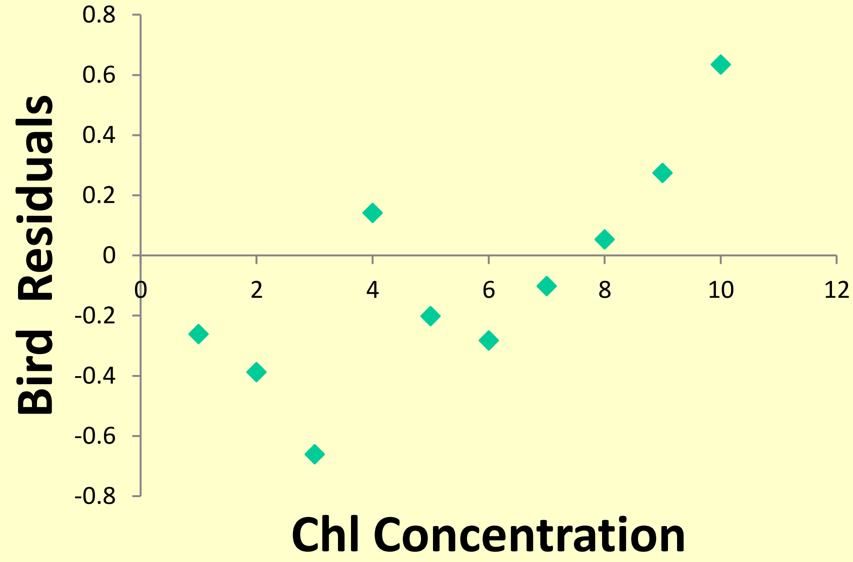
Linear Regression

How to "remove" the effect of THERM on BIRDS,
to assess the correlation between BIRDS and CHL?

Removing the Thermocline Effect



Assessing the Chl Effect

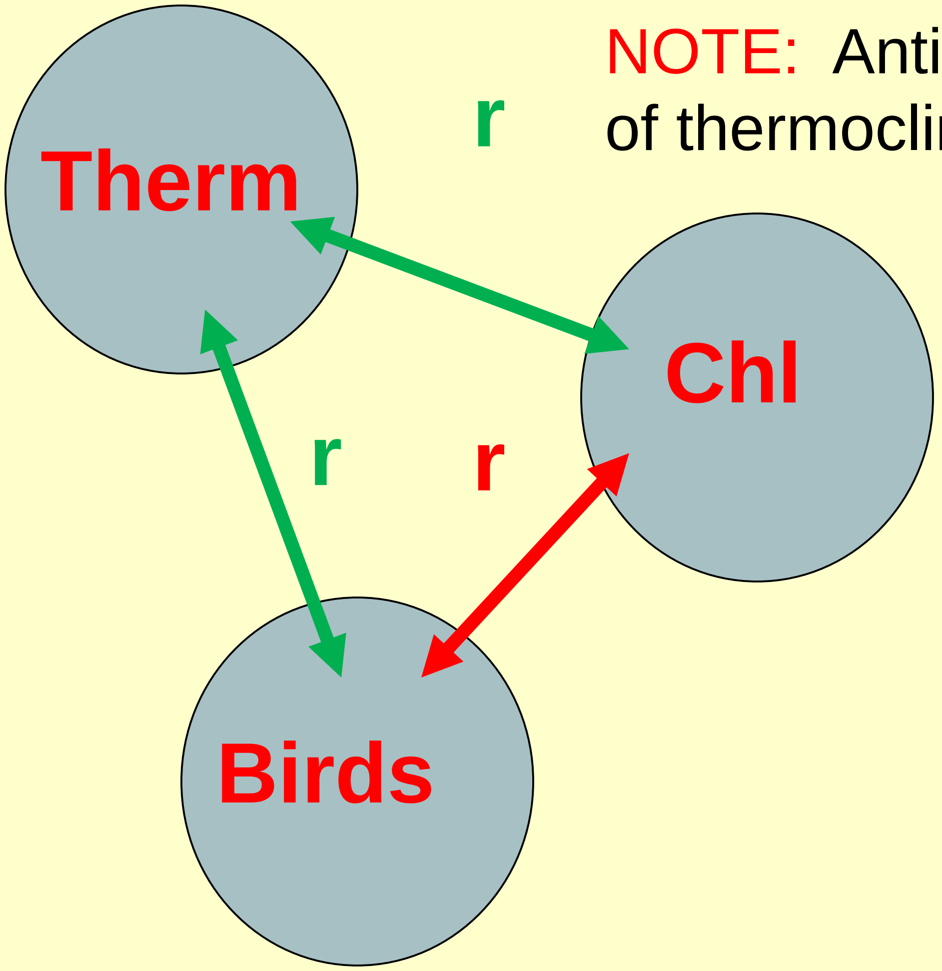


PROBLEM: This approach does not remove correlation between THERM and CHL… inherent in the CHL dataset

Partial Correlation – Using Regression

How to "remove" the effect of THERM on CHL and Birds, to assess correlation between CHL and Birds?

NOTE: Anticipate direct relationship of thermocline depth with Chl and Birds



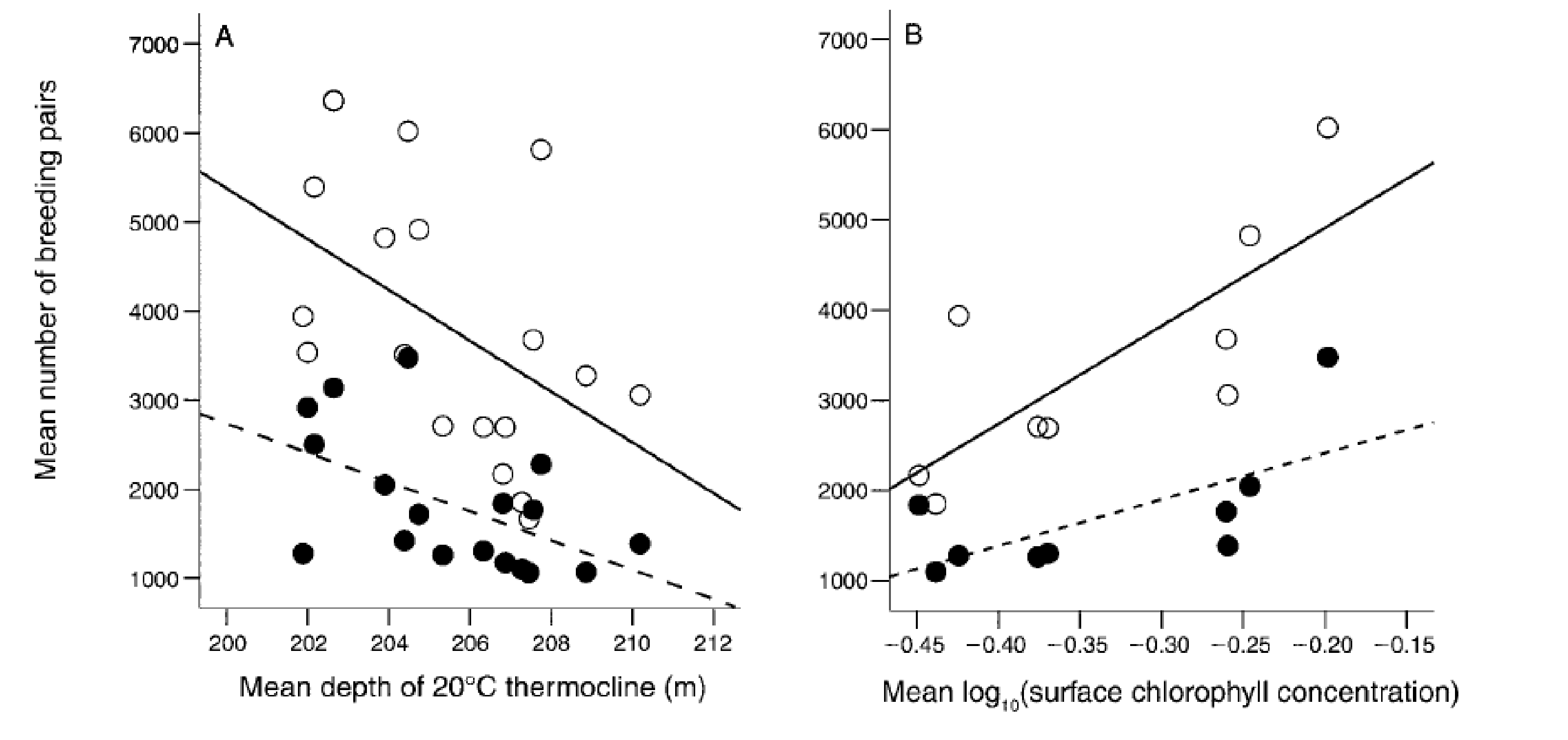
Regress Chl on Therm

Regress Birds on Therm

To remove Therm Influence, Regress:

Birds X Therm residuals on Chl X Therm residuals

Simple Regression Approach



Number of breeding pairs of pelagic terns relative to:
(A) mean annual depth of the 20 deg. C thermocline (m) and
(B) mean annual log10(surface chl a concentration) (mg/m3)
for Sooty Terns (open circles; r = 0.484, P = 0.042) and
Common Noddies (solid circles; r = 0.537, P = 0.022).

(Devney et al. 2009)

Partial Correlation – Using Regression

TABLE 2. Step-wise multiple regression analysis of breeding participation in Sooty Terns and Common Noddies vs. the two environmental predictor variables.

			r_p			
Step	Cumulative r	Increase in r	Chlorophyll	Thermocline	Model $F_{2,6}$	P
Sooty Tern						
1	**0.773**		0.773		**10.400**	**0.015**
2	**0.972**	0.199	0.962	−0.928	**50.767**	**<0.001**
Common Noddy						
1	0.737		0.701	−0.316	3.571	0.095

Notes: Variables are chlorophyll ($\log_{10}$[surface chlorophyll *a* concentration], measured in mg/m^3) and thermocline (thermocline depth at 13–16° S and 146–149° E). The statistic r_p is the partial correlation of each variable with the number of breeding pairs. Significant relationships ($P \leq 0.05$) are shown in boldface.

Conclusion:

Sooty Tern responds to both Chl and Therm

Common Noddy does not show a response

(Devney et al. 2009)

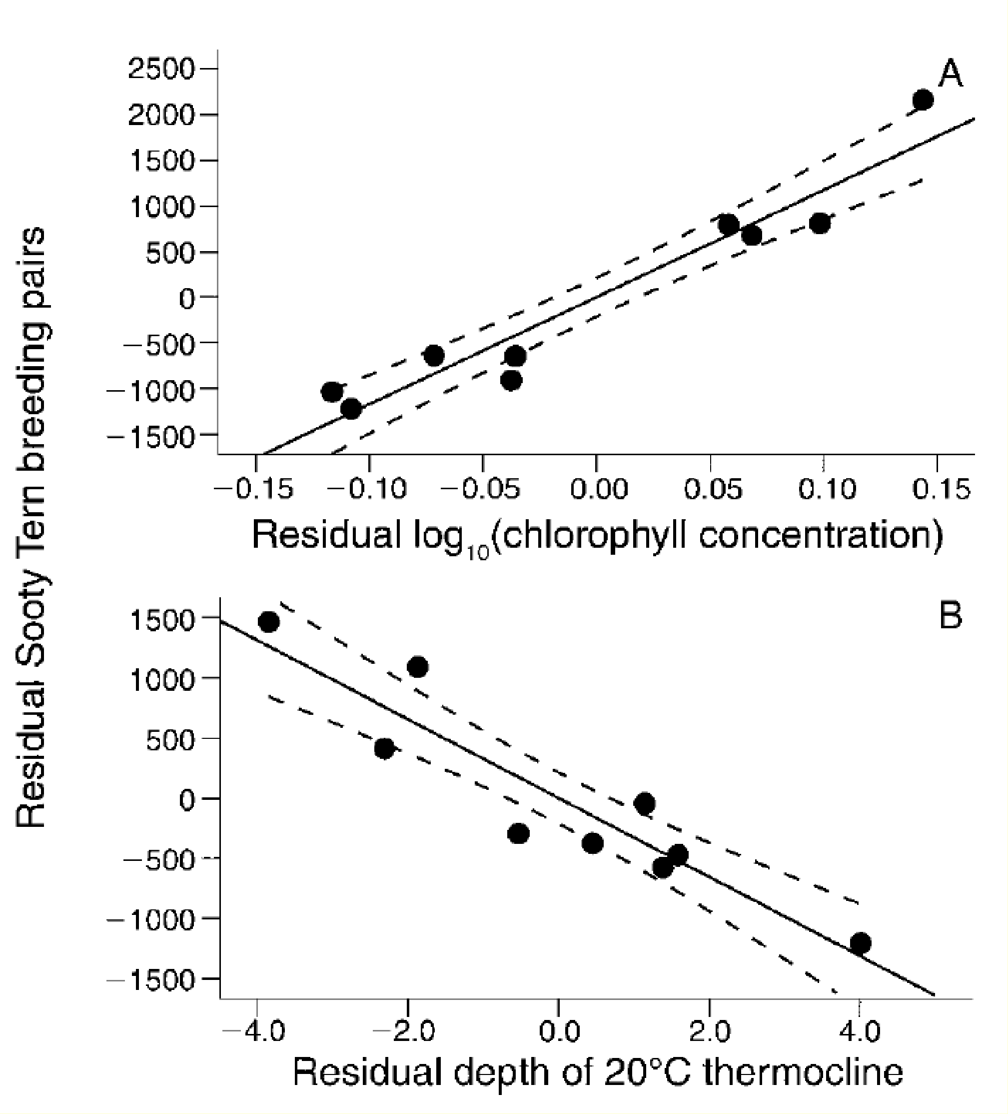
Partial Correlation – Using Regression

Partial regression of factors affecting mean annual number of Sooty Tern breeding pairs. Values are residuals from a regression of breeding pairs against the residuals of

(A) Log10 (chl a)
(r = 0.959, P < 0.001)

(B) depth of 20oC thermocline
(r = 0.925, P < 0.001).

The best-fit regression model includes both variables: chl and thermocline depth
(r = 0.972, P < 0.001).



Partial Regression – Example

Vol. 364: 15–29, 2008
doi: 10.3354/meps07486

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 29

Untangling the relationships among climate, prey and top predators in an ocean ecosystem

B. K. Wells[1,2,*], J. C. Field[1], J. A. Thayer[3], C. B. Grimes[1], S. J. Bograd[4], W. J. Sydeman[5], F. B. Schwing[4], R. Hewitt[6]

[1]Fisheries Ecology Division, NOAA Fisheries, 110 Shaffer Road, Santa Cruz, California 95060, USA
[2]Long Marine Laboratory, University of California, Santa Cruz, 100 Shaffer Road, Santa Cruz, California 95060, USA
[3]PRBO Conservation Science, 3820 Cypress Drive #11, Petaluma, California 94954, USA
[4]Environmental Research Division, NOAA Fisheries, 1352 Lighthouse Avenue, Pacific Grove, California 93950, USA
[5]Farallon Institute for Advanced Ecosystem Research, PO Box 750756, Petaluma, California 94975, USA
[6]Southwest Fisheries Science Center, NOAA Fisheries, 8604 La Jolla Shores Drive, La Jolla, California 92037, USA

- Rationale: Relate marine predator responses to local, regional, and basin-wide oceanographic variables.

(Wells et al. 2008)

Partial Regression – Example

- Analyses: Used path analysis to identify correlations between biotic and environmental variables, determine direct and indirect relationships of variables and visualize likely mechanisms forcing productivity of 3 trophic levels in CCS.

PROS:

Identify correlations between variables
Each variable can have both direct and indirect effects

CONS:

The structure of path models was determined a priori based on understanding of oceanographic and trophic interactions

(Wells et al. 2008)

Partial Regression – Example

Physical and biological variables:

from wind to krill abundance to seabird productivity

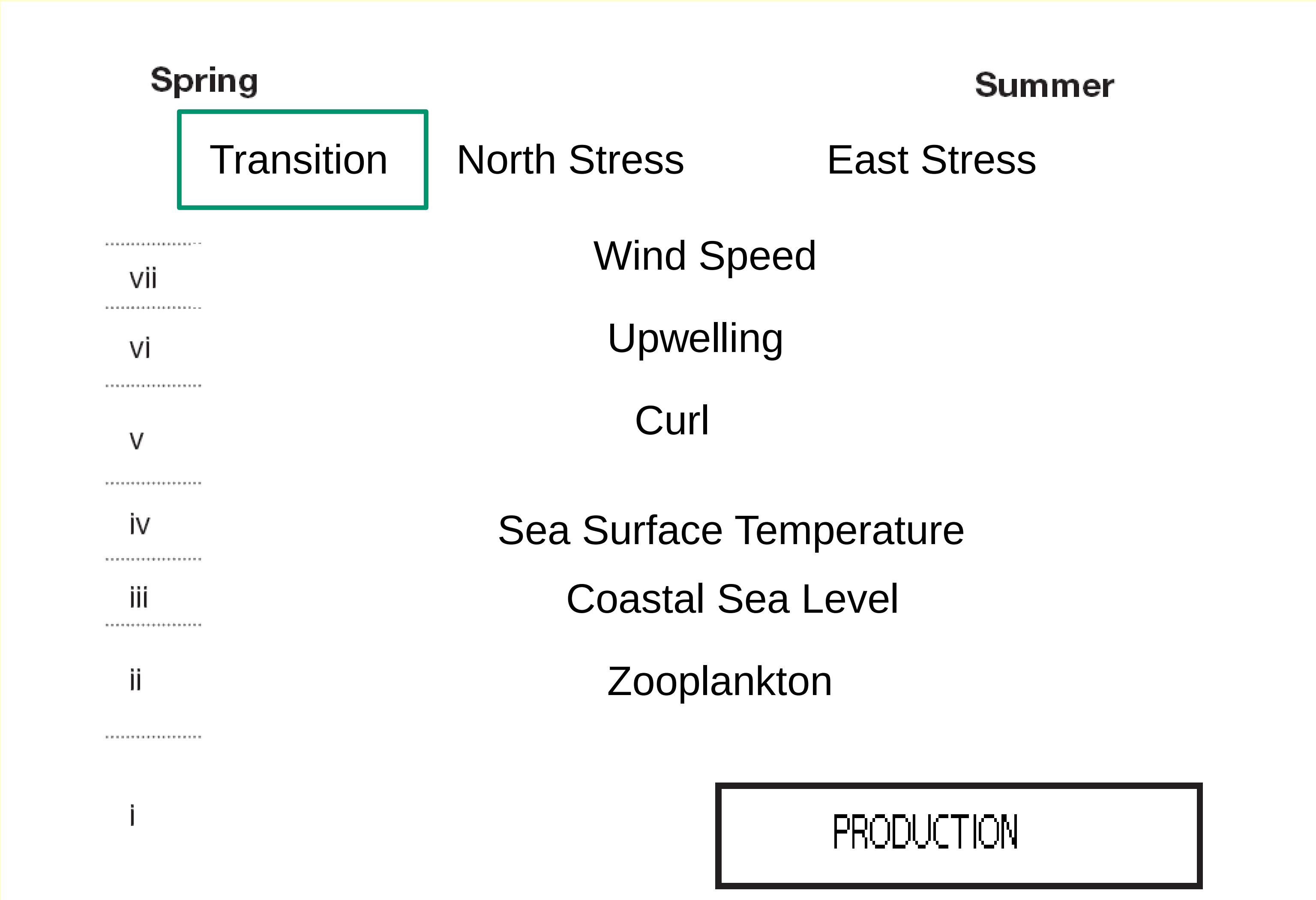
Table 1. Eight environmental variables used to characterize the ocean environment of central California (Fig. 1) were collected from various sources for the years 1972 to 2005. Variables were averaged within the region and into spring (March, April and May) and summer (June, July and August) values. These seasonal values (15 variables in total) were included in season-specific path analyses, and those variables with effects were included in partial least squares regression analyses to model variation in productions. The variables, identifiers used in figures, their sources and associated noteworthy comments are listed

Variable description	Variable identifier	Source
Transition date[a]	Transition	Schwing et al. (2006)
Easterly pseudo-wind stress[b]	East stress	Worley et al. (2005)[c]
Northerly pseudo-wind stress[b]	North stress	Worley et al. (2005)[c]
Non-directional wind speed; scalar wind speed	Wind speed	Worley et al. (2005)[c]
Upwelling	Upwelling	Pacific Fisheries Environmental Laboratory station[d]
Curl[e]	Curl	Pacific Fisheries Environmental Laboratory station[d]
Sea surface temperature	SST	Worley et al. (2005)[d]
Coastal sea level height	CSL	University of Hawai'i Sea Level Center station, San Francisco, CA

[a]Used only for spring
[b]Converted to meteorological notation
[c]International Comprehensive Ocean Atmosphere Data Set using 1° resolution averages
[d]39° N, 125° W (see Fig. 1)
[e]Spring values in 1978 were unavailable

(Wells et al. 2008)

Partial Regression – Example



(Wells et al. 2008)

Partial Regression – Example

The first stepwise regression performed with the production variable (level-I) as the dependent variable and the other (higher) variables as independent (explanatory) variables.

The variable nearest to the production variable in the structural model that was significantly related to the production variable, was linked to the production variable in the causal model.

This variable was then treated as the dependent variable, with all variables above it in the model structure as independent.

The procedure was repeatedly performed until only exogenous (highest level) variables remained. Finally, evaluated the correlations among exogenous variables. **(Wells et al. 2008)**

Partial Regression – Example

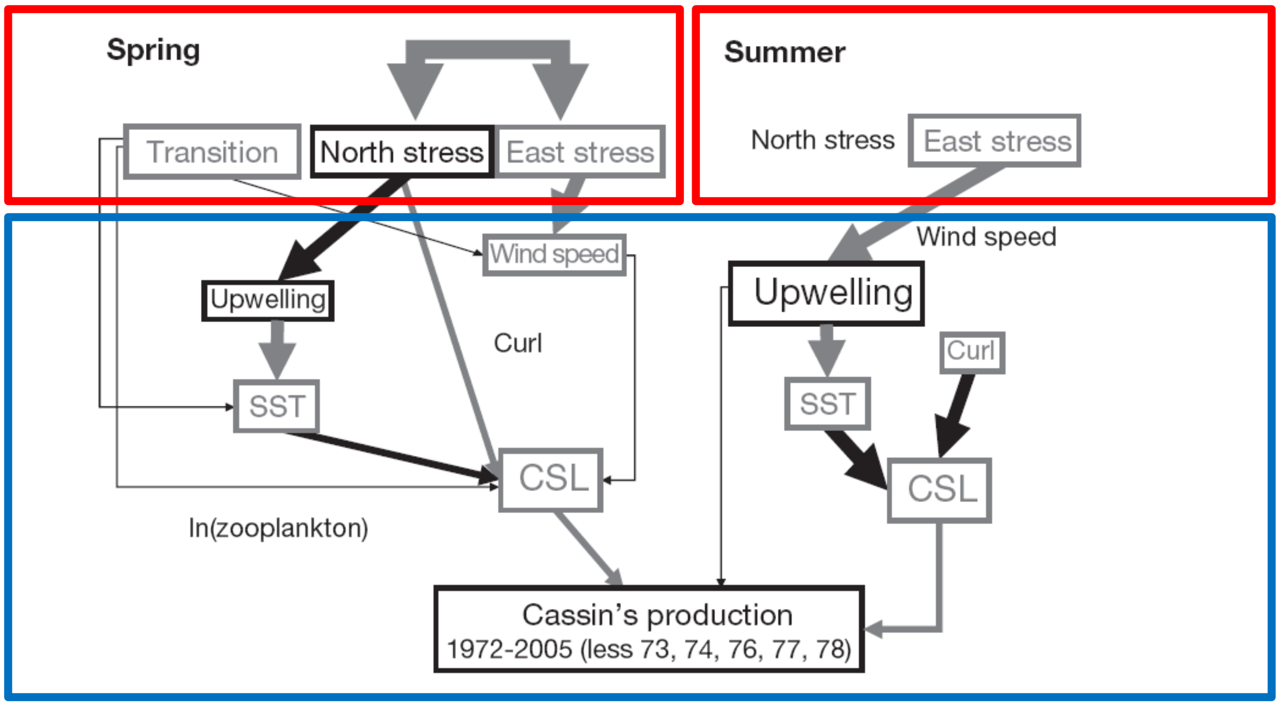
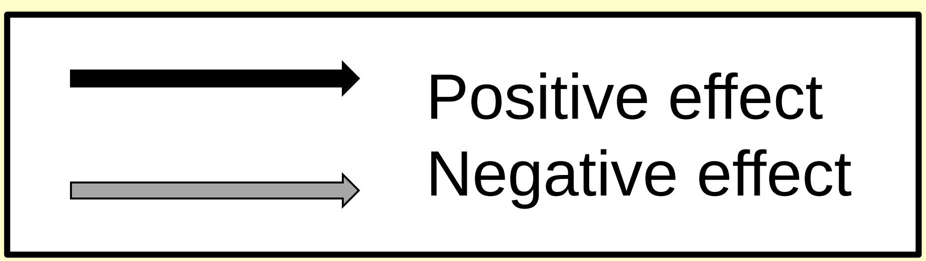


Fig. 8. Zooplankton and *Ptychoramphus aleuticus*. Spring and summer path analyses on Cassin's auklet. Legend for Fig. 2 is appropriate here except there is an additional arrow line weight to represent a standardized slope coefficient of 0.81 to 1.00

Correlation between "external drivers"

Partial regression with "internal variables"



(Wells et al. 2008)

Partial Correlation – Summary

Order of correlation:

A "first order partial correlation" has a single control variable. A "second order partial correlation" has two control variables … etc.

A "zero-order correlation" is one with no controls: it is a simple correlation coefficient.

Un-partialed (regular) correlations are *zero order*.

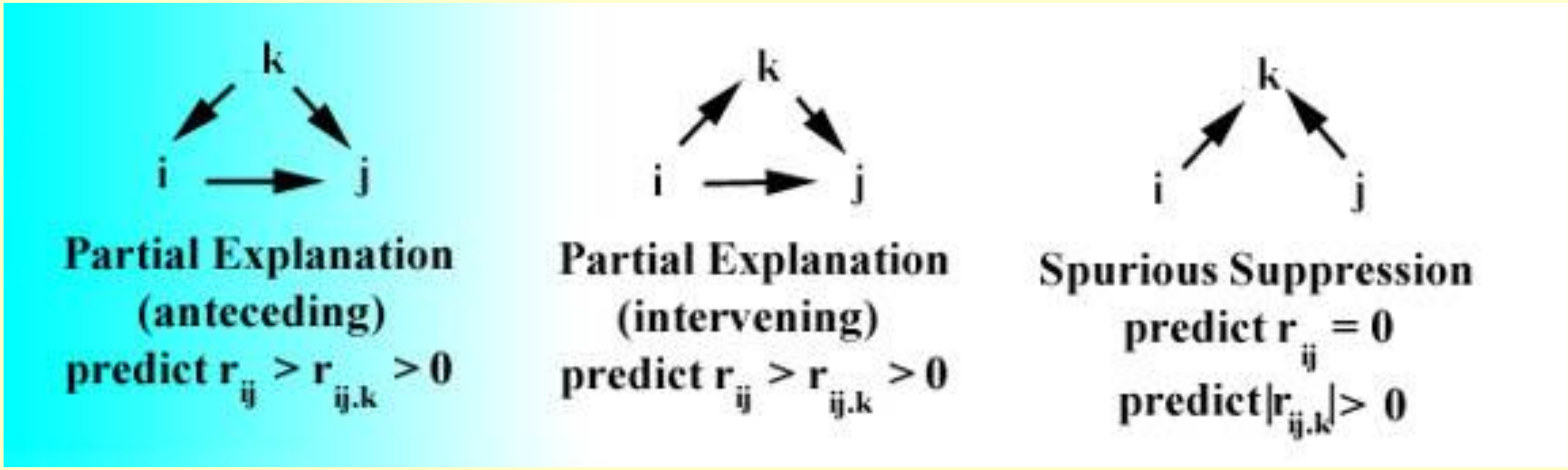
If we partial one variable out of a correlation (e.g., r12.3), that partial correlation is a *first order partial correlation*.

If we partial out 2 variables from that correlation (e.g., r12.34), we have a *second order partial correlation*

… and so forth.

Partial Correlation – Summary

- Three generic types of third variable influences:



Anteceding Variable

Intervening Variable

Spurious Suppression

Partial Correlation – Summary

- We can use a formula to compute first order partials, or we can use simple regression to compute the residuals – which are correlated.

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

- For example, to compute r12.3, we regress X1 and X2 on X3 and then compute the correlation between the residuals.

- This approach would compute the correlation between X1 and X2, controlling for the influence of X3 on both.

Partial Regression – Summary

- We can also use a formula to compute second and higher order partials, or we can use multiple regression to compute the residuals – which are then correlated.

- For example, to compute r12.34, we could regress each of X1 and X2 on both X3 and X4 simultaneously and then compute the correlation between the residuals.

- This approach would compute the correlation between X1 and X2, controlling for the influence of both X3 and X4.

The Big Picture - Summary

- Partial correlation and partial regression approaches allow us to deal with collinearity between multiple variables

- However, these approaches are restricted to a small number of explanatory variables – for logistical reasons

- Furthermore, these approaches require a pre-conceived notion of the cause – effect relationships (hierarchical model)

Implications for Multi-variate Statistics:

- Assess zero-order correlations of explanatory variables

- This data exploration provides insights into why multivariate methods combined these cross-correlated variables

Homework I – Due Feb 8

- *Objectives:*

Showcase quantification of collinearity

Use partial correlation analysis

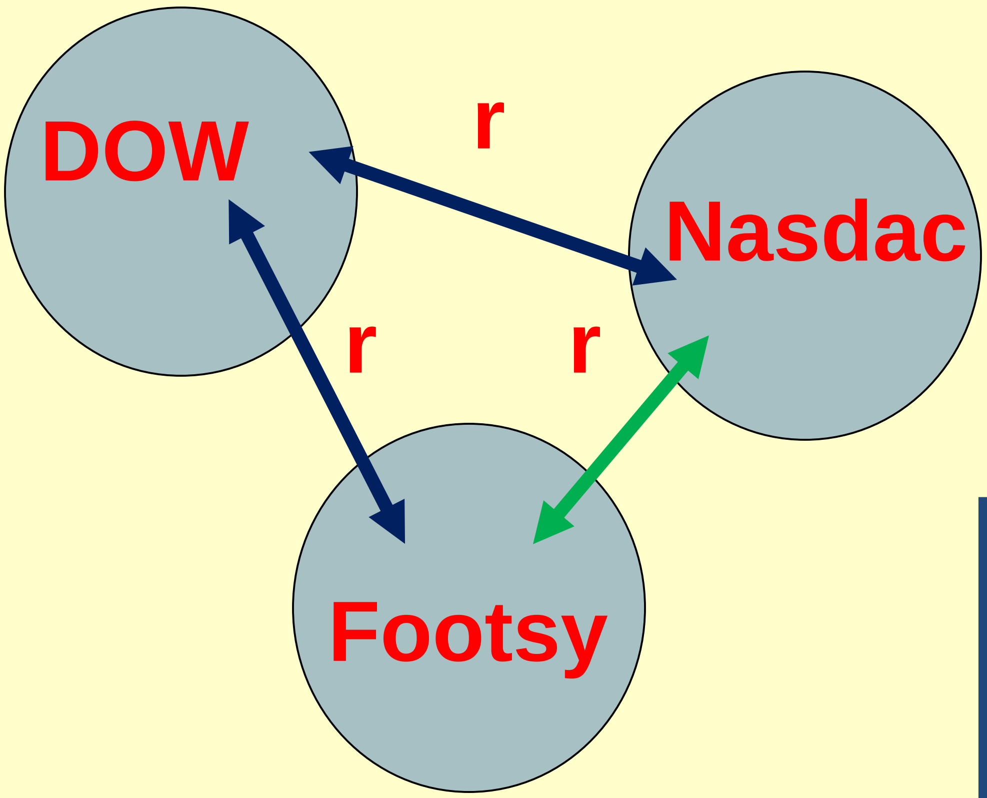
Use partial regression analysis

- *NOTE:*

I will not discuss homeworks the day they are due

Partial Correlation - Stocks

Quantify shared variability of 3 stock exchange metrics:



NOTE:
Anticipate influence of global DOW index on the other two regional indices

Method 1:

r (DOW,Nasdac . Footsy)

Method 2:

Regress Footsy on DOW

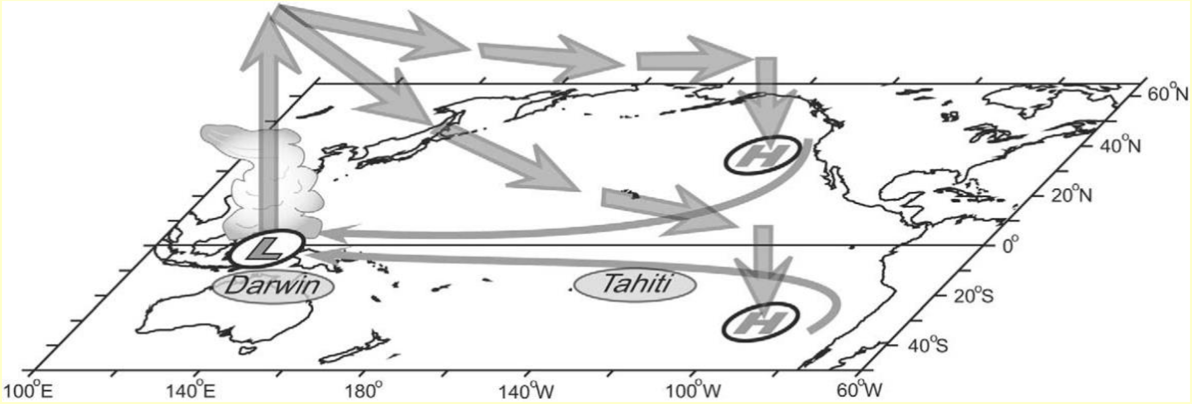
Regress Nasdac on DOW

Regress Footsy Residuals on Nasdac Residuals

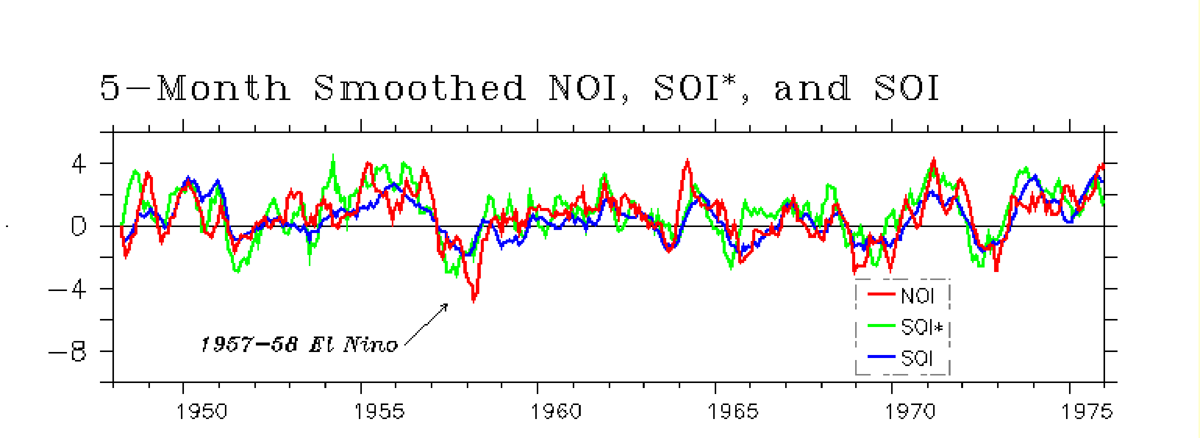
Partial Correlation – Oceanography

- Partial approaches used in causal models (hypotheses).
- Also useful to quantify co-varying patterns (exploratory) .

Correlated "spatial footprints"



Co-occurring patterns (indices)



(Schwing et al. 2002)