Intro to Analysis of Multi-variable Data: 12-hour course, 1995

Outline

Session 1

- Y — Scales
  Summary Statistics / Parameters
- X & Y — Various Configurations
  Measures of relationship Y <--> X
- X_1, X_2 & Y — Objectives of M.V.A.
  - Fairer Comparisons
  - Sharper Comparisons
  - "Determinants of ..."
  - Prediction
  - Effect modification
    ("Different Slopes for Different Folks")
- "Non-regression" Methods
  - Risk/Rate Differences/Ratios
  - Differences in Means
  - Slopes

Session 2

- Multiple Regression
- Adjustment / Noise-reduction :- Means
- Adjustment:- Proportions

Session 3

- Several determinants
- Collinearity
- Effect Modification / Interaction

Session 4

- Computing
- Case Studies

Session 1:

Outline

- Y — Scales
  Summary Statistics / Parameters
- X & Y — Various Configurations / Displays
  Measures of relationship Y <--> X
  references: M&M Ch 2
- X_1, X_2 & Y — Roles of X_1 & X_2 :
  - Fairer Comparison of levels of X_1
    BIAS REDUCTION --> X_2 is a "Confounder"
  - Sharper Comparison of levels of X_1
    MORE PRECISION --> X_2 not necessarily confounder
    but produces considerable addnl. variation in Y
  - Interest in Both X_1, X_2 as determinants of Y
    X_1 and X_2 have same SYMMETRICAL status
  - X_2 "modifies" relationship between X_1 and Y
    DIFFERENT Y<->X_1 relationship
    for DIFFERENT levels (subgroups) of X_2

Examples of (Y, X_1, X_2 ...) Data

- Admissions of Males & Females to Berkeley Graduate Schools
  - overall and faculty by faculty
- Birthweight - Gestational Age ; Gender
- Fatalities & Speed Limit Change - Time
- Low Birthweight - Alcohol ; Smoking ; Social Class
- Intelligence Quotient (IQ) - Mother's Milk; Other Variables
- Stature(height) of Children on Tetracycline -
- Lung Function of Vanadium Factory Workers
  - vs. reference group (matched for smoking and age)
    that was 3.4 cm different in average height
- Blood Pressure and Altitude - age; height; weight; country
- Weight - Age ; Social Class
- longevity - sexual Activity; Size

X_1, X_2 & Y:

If primary interest is in X_1 contrast, and X_2 is either a "Confounder" or produces considerable additional variation in Y that acts as 'noise'.

Simplest case: X_1 is measured on a 2-point scale (binary) so compare Y in those with $X_1 = 0$ vs. in those with $X_1 = 1$;

NON-REGRESSION METHODS

Paired / Less Finely Stratified Observations (X_2 : pair / stratum)

X_2	$X_1 = 0$	$X_1 = 1$	ΔResponse *
1	(ave.) response	(ave.) response	d
2	(ave.) response	(ave.) response	d
..	…	…	…
..	(ave.) response	(ave.) response	d
$\sum$			$\frac{\sum w \bullet d}{\sum w}$

* using d generically to represent any comparison
(could be difference, ratio, etc...)

Key: (Weighted) Average of "Within-stratum"
or "other-factors-being-equal" comparisons.

Confounding:

Δ of aggregated responses NOT SAME AS aggregate of Δ's

References:

counted and measured Y's: Smith & Morrow, §14.6
AAHOVW
Miettinen §11-16

counted Y's: Walker §8 & 13
KKM § 13

Collinearity

Example of the issue: Suppose that in a study of workers aged 45-65 to quantify the degree to which hearing loss was affected by their exposure to the noise from heavy machinery, the number of years of exposure to this noise and the extent of hearing loss were determined for each person. A multiple regression is planned to assess the effect and to take the person's age into account (hearing loss generally becomes worse with age, even if there is no unusual occupational exposure).

What is the correlation between age and cumulated exposure likely to be?

If it is very high, what will it do to the estimate of the regression slope of loss on exposure?.

If it is low, what will it do? If you think it will do very little, would you bother to include age in the regression? [This question has to do with reduction of noise and making comparisons sharper].

If you had a choice of which workers to select from a larger available group, would you choose on a purely random basis, or on some other basis? Why?

See some examples on next page. The panel on the extreme left shows the distribution of age and exposure (both in years), with a fairly strong positive correlation. An example of a 'stratified sample' is given next to it (upper panel). Here the selection is constrained to obtain persons equally from all 4 quadrants. This makes it easier to separate the effect of age from the effect of exposure. An example of an 'unstratified sample' is given in the lower panel. Here the selection is simply a 'miniature' of the parent distribution and so there will be greater difficulty in separating the effect of age from the effect of exposure.

Suppose that in fact the mean hearing loss for persons of a certain age and exposure is as follows:

$$\text{mean} = 0.3 \bullet (\text{age} - 25) + 0.4 \bullet \text{exposure}$$

and that the inter-individual variation around this mean is Gaussian with a SD of 3. In technical language, we say that ß[exposure] = 0.4 and that ß[age] = 0.3, and that the SD of the 'residuals' is 3.0.

On the right hand side of the following page the effects of the collinearity on our estimates of the two ß's are displayed in list and graphic mode for 10 unconstrained and 10 constrained (stratified) random samples. The message from these is that the estimates of the ß associated with exposure are more variable (and so less dependable) when the samples have collinearity. (the same is true for the estimates of the ß for age).
In the extreme, if the collinearity between age and exposure were close to a correlation of 1, the estimates of the ß for exposure could oscillate even more, and could go from being quite negative to quite positive. The only thing that would remain reasonably stable is the sum of the estimate of ß for exposure and of the ß for age (i.e. the sum of the two estimates would be close to 0.4 + 0.3 = 0.7, but an equation with the estimate of ß[exposure] = –1.2 and ß[age] = +1.9 {or for that matter ß[exposure] = +2.3 and ß[age] = –1.6} would do an equally good job of predicting the responses (all the individuals would be spread out along the diagonal in the age vs. exposure diagram). You can see some of this compensatory behaviour of the two estimates in the plot in the panel on the right (estimates from "unstratified" samples), where there is a strong negative correlation between the two estimates.

In the previous example, if females, because of their longer hair or greater tendency to wear ear-protectors, or because of some biological factor that might make them less susceptiple to noise-induced hearing loss, were analyzed separately from males, how would the regression coefficients for hearing loss on years of exposure compare in the two sexes?

Effect Modification = "Different Slopes for Different Folks"

Can we combine the separate equations for males and females into one?

A similar example of combining two equations into one: How to estimate ideal body weight (based on findings of a Harvard study)

For Women: 100 pounds for a height of 5 feet, with five additional pounds for each added inch of height

For Men: 110 pounds for a height of 5 feet, and six additional pounds for every added inch of height

Since 5 feet = 60 inches, and letting H = height in inches – 60, the equations become:

Women: weight = 100 + 5•H
Men: weight = 110 + 6•H

If denote Women by a variable G(ender)=0 and Men by G=1, we can combine the 2 equations

weight = 100 + 10•G + 5•H + 1•G•H

Terminology: Note that the use of the product G•H as an additional variable in the regression equation is called an 'interaction' term. If the coefficient associated with this variable were 0, we would have 'no statistical interaction' (i.e. we would have the 'same slope for different folks'.

Thus the ideas of 'effect modification' and 'statistical interaction' are really the same: epidemiologists tend to use the former and statisticians the latter.

The trouble with the word interaction is that it refers to a purely numerical trick to write the equations for 2 or more non-parallel lines in a single compact equation. Unfortunately, users of the equations sometimes try to give the word a biological meaning. But by suitable transformations, one can sometimes transform non-parallel curves into parallel lines and vice versa, so any 'interaction' term has to be viewed in the context of the scale used.

Preamble / Motivation / ...

- Easy to carry out (just click!)
- Easy to be "glib" about what it accomplishes
- BUT ... WHY use it ??? HOW to explain to father-in-law?

• ***If interested in separate contributions of each of several variables...***

are there any situations where one can assess them one at a time? *i.e.*

assess a particular X while ignoring the others ...
assess a different X while ignoring the others ... ?

or does one have to assess them simultaneously ?

• ***If interested in ("net") contribution of ONE particular variable...***

are there situations where one can assess it while ignoring the others ... ?

or does one always have consider the other X's as well ?

Answers ... Illustrated by examples

- birthweight as function of gestational age and gender
- weight in relation to age and height
- breast milk and subsequent IQ in children born preterm
- increase in heating costs after adding a room to a house
- decrease in longevity if greater amount of sexual activity

Multiple Regression Equation

$$Y_{X1\ X2\ \ldots} = \mu_{Y\ |\ X1\ X2\ \ldots} + \varepsilon$$

$$\mu_{Y\ |\ X1\ X2\ \ldots} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

How to describe it ...

in words / symbols

$\mu_{Y\ |\ X1\ X2}$ .. as a function of X1 X2 ..
(don't forget the ε's with SD σ about μ's)

geometrically

"plane" or "surface" of means

(in case of 2 X's) without leaving "2-D"

as contour map (cf web page)

using links to simpler procedure..

as a sequence of simple linear regressions
(but be careful: see my notes on Ch 2/9 of M&M)

Meaning of β_i

$$\beta_i = \frac{\delta \mu_{Y \mid X1 \; X2 \ldots}}{\delta X_i}$$

difference in μ_Y for a 1 unit difference in X_i but no difference in other X's, i.e. all other X's held "constant"

Main Purposes

- Summarization / Description
- Adjustment (Bias Reduction)
- Increased Precision of estimates of specific β_i's
  (by removing extraneous variation)
- Prediction
- Interpolation / Smoothing
  "borrowing strength" (e.g. estimates of outcome of prostate cancer if sparse data in some age-histologic grade "cells")
- Polynomial Regression
  (several powers of 1 X -- each power is a *term* in regression; can also have other X's in equation)

Assumptions

see G&S page 54 [page 58 in 2nd ed] ;
see also comments in my notes on ch. 3

Parameters	**Estimates of these** *(by computer!)*
1. β_0	$b_0 \pm t\, SE[b_0]$ *[β_0 seldom of interest]*
2. β_i	$b_i \pm t\, SE[b_i]$
3. $\sigma_{Y \mid X1\ X2 \ldots}$	$\sqrt{\frac{\Sigma (y_i - [b_0 + b_1 x_i + b_2 x_2 \ldots])^2}{n - \#\text{ of b's fitted}}}$ *("Root Mean Squared Error")*
4. $\mu_{Y \mid X1\ X2 \ldots}$	$b_0 + b_1 X_1 + b_2 X_2 + \ldots$ $\pm t\, SE[\text{ thereof }]$
5. $Y_{\mid X1\ X2 \ldots}$	$b_0 + b_1 X_1 + b_2 X_2 + \ldots$ $\pm t\, SE[b_0 + b_1 X_1 + b_2 X_2 + \ldots + \varepsilon]$ (Interval for **Y|X wider** than for $\mu_{Y \mid X}$)

Multiple Correlation Coefficient

- a helpful way to look at least squares estimate (scalar)

$R_{Y \text{ and best linear combination of X's}}$

Confounding: Reducing it by Regression

Preamble

- Don't overlook classical, "non-regression" methods
- Regression methods are more "synthetic" (i.e. "artificial")
- Cf chapter 3 by Anderson et al. (c622; readings from aahovw)

Definitions ... / synonyms

Original (statistical, in design of experiments)

- inability to estimate higher order interactions
  (so typically assume they are zero)

- "mixed up with other effects" or "inextricable"

Epidemiological

- (osm)

Other terms

- "Lurking" (i.e. "hidden") variable

- "Simpson's Paradox" is the most extreme form

*(see collection of Simpson's paradox examples under **Other Resources** on **c626**)*

Examples...

- Does using a Macintosh lead to sloppier writing? **a**
- Better Service from Canada Post after "Major Restructuring"**a**
- Salaries of Master's and PhD's **a**
- Outcomes of Pregnancy during Residency for women and wives of their male classmates• Admissions of Males & Females to Berkeley Graduate Schools **b**
- Percentage of White & Black Convicts Receiving Death Penalty **a**
- Intelligence Quotient (IQ) - Mother's Milk; Other Variables **a**
- Lung Function of Vanadium Factory Workers **Other resources, c697**
  - vs. reference group (matched for smoking and age) that was 3.4 cm different in ave. height
- Blood Pressure and Altitude - age; height; weight; country **b**
- Longevity - Sexual Activity; thorax size **c622**
- Fatalities & Speed Limit Change - Time **a**
- NEURODEVELOPMENT OF CHILDREN EXPOSED IN UTERO TO ANTIDEPRESSANT DRUGS **b**
- What Does It Take to Heat a New Room? **dataset, c697**

---

a notes on Ch 2, c607 **b** resources this course (678), session 5

Preamble

- Should not be in same chapter with confounding...

- a very different topic !! (can have both, but ... see diagram)

Definitions ...

Interaction (statistical)

- "Non-additivity" of "effects" in regression

- need for product term in regression analysis (osm)

- scale dependent

(Effect) Modification (epidemiological)

- Inconstancy of a parameter of a relation over other subject characteristic (osm)

- Different slopes for different folks (jh)

"Modifier (of a relation)

- A characteristic (of individuals) on which a parameter of a relation depends (osm)

Examples...

• Equation for Ideal Weight as function of Height

- modification by Gender

• Average Earnings as function of Education / Age

- modification by Gender

• Decline in Bone Density with Age

- Different in 19th and 20th Centuries

• ?Can hit further with aluminum than wood baseball bat?

- Difference depends on where on bat one hits ball

• Changes over time in injury rates

- Different in intervention and reference areas?

Translating these into regression equations ...

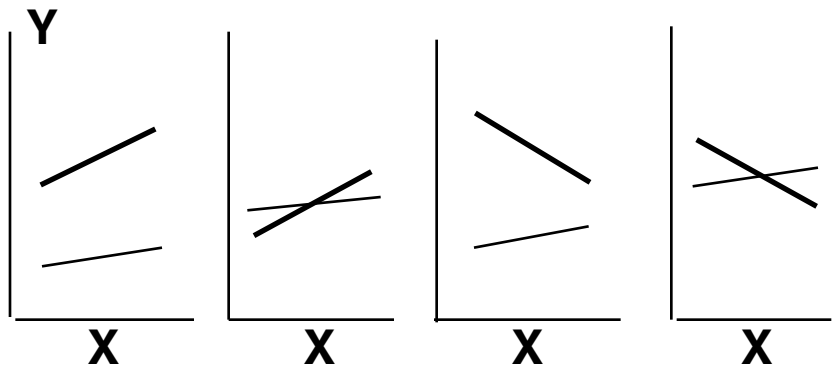
- relation between Y and X

- "modifier" variable M

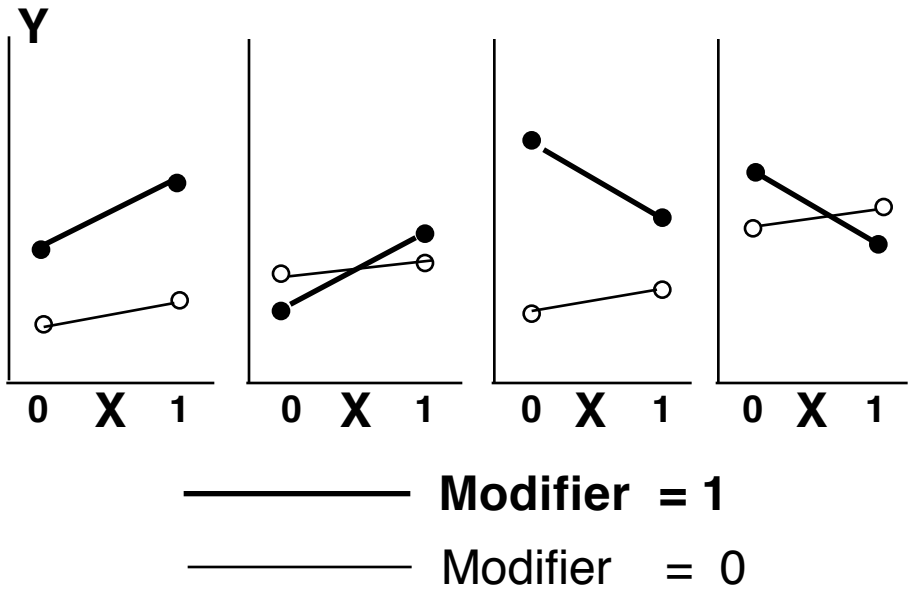
$$E[Y \mid X, M] = B0 + B1.X + B2.M + B3.(M.X)$$

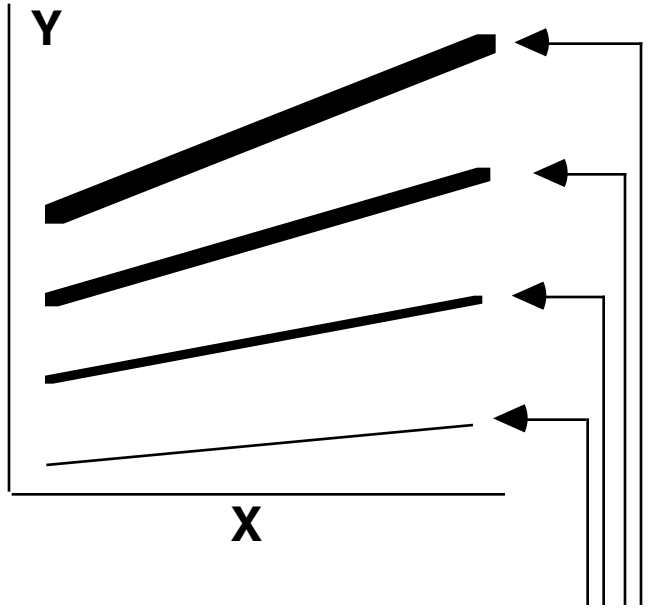
- Special cases..

X continuous, M binary



X binary, M Binary

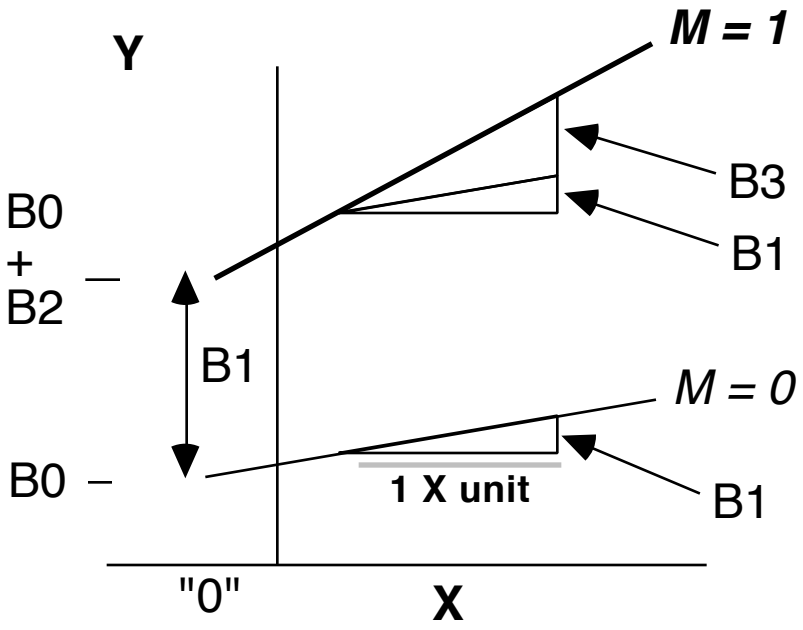




Quantitative levels of Modifier M

Meaning of the coefficients

X continuous, M Binary



- helpful ways of rewriting the equation

E[Y | X, M] = B0 + B2.M + (B1 + B3.M).X

Special issues

• mathematical symmetry of equation

E[Y | X1, X2] = B0 + B1.X1 + B2.X2 + B3.(X1.X2)

= B0 + B2.X2 + (B1 + B3.X2).X1

X2 modifies the Y<->X1 relation

= B0 + B1.X1 + (B2 + B3.X1).X2

X1 modifies the Y<->X2 relation

• to a regression program, X1.X2 product terms are just like any other terms.. but they tend to be correlated (collinear) with the components from which they are made, so...

*** user should "center" the components before *** making (or having computer make) products ***

(will see example in injury prevention study)

Translating equations back into lines ...

- **If M is binary...**

start with the M=0 case

$$B0 + B1.X + B2.M + B3.(M.X)$$

$$= B0 + B1.X + B2.0 + B3.(0.X)$$

$$= B0 + B1.X$$

===> straight line in X with intercept B0 and slope B1

"turn on" the M=1 toggle...

$$B0 + B1.X + B2.M + B3.(M.X)$$

$$= B0 + B1.X + B2.1 + B3.(1.X)$$

$$= B0 + B1.X + B2 + B3.X$$

collect terms that do not involve X & those that do..

$$(B0 + B2) + (B1 + B3).X$$

===> straight line in X with intercept (B0 + B2) and slope (B1 + B3)

- **If M is continuous...** as above with several M values