Epidemiology VHM 811 Course
Fall 2009, Atlantic Veterinary College, PEI
Henrik Stryhn

Index of Lecture 10a

Page	Title
1	Practical information
2	Things you should know about now...
3	Confounding — introductory remarks
4	Confounding — demonstration example
5	Conditions for confounding
6	Confounding - real example
7	Counterfactual definition of confounding
8	Working definition of confounding
9	Non-collapsibility of odds-ratios
10	Control of confounding — overview
11	Graphical analysis for confounding
12	Example of multifactorial graph
13	Illustration: association created by control
14	Mantel-Haenszel procedures
15	Feedlot data — confounding by province
16	Feedlot data — M-H analysis for IBR
17	Stata commands for M-H analysis
18	Confounding — other situations
19	Stata do-file

Practical Information

Welcome to the sessions on Confounding and Interaction!

Schedule
— refer to webpage for information, in particular:

- lecture handouts,
- datasets (VER + extras),
- homework for next session,
- sample problems (home assignment and exam) from previous years.

Today's lecture:

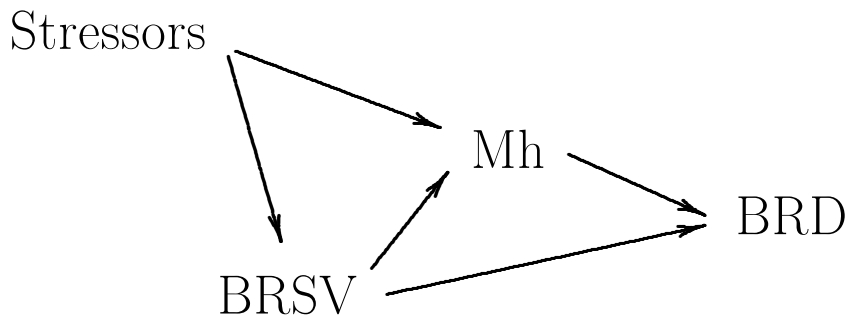
- confounding — arguably one of the key concepts in epidemiology,
  - examples and definitions,
  - procedures for detecting/dealing with confounding,
  - brief Stata demonstrations (but see do-file),
  - some slides may be postponed until tomorrow,
- bear with me (and ask!) if I assume knowledge you do not quite have.

Home work for Friday: VER Problem 13.7,

- go through both using calculator and Stata,
- dataset via webpage link (not included with VER),
- don't go into details about interaction (question (c)).

Things You Should Know About Now ...

Causal diagrams, e.g. for BRD problem (Sections 1.6,1.9)[1]



* nodes $\sim$ variables, and arrows $\sim$ causal relations,
* horizontal axis $\sim$ time (right=most recent),
* intervening (intermediate, mediating) variable: on causal pathway between two nodes.

Measures of association/effect for binary outcome (disease), computed from estimated probabilities $\hat{p}_1$ and $\hat{p}_0$ in two groups:

- risk difference (RD) $= \hat{p}_1 - \hat{p}_0$,
- relative risk, or risk ratio, (RR) $= \hat{p}_1/\hat{p}_0$,
- odds-ratio (OR) $= [\hat{p}_1/(1-\hat{p}_1)] \,/\, [\hat{p}_0/(1-\hat{p}_0)]$.

Epidemiological designs:

- cross-sectional study,
- cohort study,
- case-control study.

[1] BRD = bovine respiratory disease; BRSV = bovine respiratory syncytial virus; Mh = Mannheimia hemolytica bacterium.

Confounding - Introductory Remarks

Confounding:

- from Latin (confundere $\sim$ to mix together),
- loosely stated means that some effects are mixed up (are not, or cannot be, distinguished).

Basic notation/terminology:

- outcome Y (of any type, but binary in our examples),
- exposure E (of any type, as above),
- extraneous factor of interest Z (measured or unmeasured; of any type, as above),
- confounder (or lurking variable) Z: extraneous factor that exerts confounding of the relation $E \rightarrow Y$,
  - note that a confounder is always tied to both outcome and exposure,
  - confounding depends on chosen outcome parameter,
  - confounding also depends on reference population.[2]

Two other usages of term confounding:

- contingency table analysis: *non-collapsibility* (L10a–9),
- experimental design (statistics): deliberate mixing of some treatment effects with block effects (*aliasing*).

[2] It is possible, but generally less interesting, to consider confounding based on an observed sample only.

Confounding — Demonstration Example

Example 13.1: constructed data on the relationship between BRD (Y), Mh (E) and BRSV (Z):

			Conditional tables on BRSV			
Disease/	total		BRSV +		BRSV −	
Exposure	Mh +	Mh −	Mh +	Mh −	Mh +	Mh −
BRD +	240	40	220	10	20	30
BRD −	6260	3460	5280	490	980	2970
risk (%)	3.69	1.14	4.00	2.00	2.00	1.00
RD (%)	2.55		2.00		1.00	
RR	3.23		2.00		2.00	
OR	3.32		2.04		2.02	

Apparently paradoxial finding:
the combined effect measures are larger than in both of the two BRSV groups; in particular,

* how can the combined risk (RR) be 3.23 when RR=2 in both groups?
* can we trust the value 3.23? (the answer is no).

Intuitive explanation:

* BRSV-positive animals are more likely to have BRD than BRSV-negative animals (proportions 230/6000=3.8%, and 50/4000=1.3%, respectively),
* Mh bacteria are far more common among BRSV-positive than BRSV-negative animals,

$\Rightarrow$ part of the (total) Mh effect is due to BRSV.

Conditions for Confounding

Definition of confounding:

- mathematically not easy; best attempt uses counterfactual arguments (L10a–7),
- literature agrees (and focuses) on *necesssary* (but not sufficient) conditions for confounding.

3 necessary conditions for Z to confound the relation $E \rightarrow Y$:

- Z must be a risk factor for Y; more precisely:
  - at the reference level of E, i.e. within "exposure-negative subjects" (because the risk must not be caused by a link with E)
- Z must be associated with E in the source population; specifically,
  - cohort study: Z and E must be associated at the start of the follow-up period,[3]
  - case-control study: Z and E must be associated in the controls,[4]
- Z must not be affected by E (which would make Z an *intervening variable* between E and Y), and Z must not be an effect of Y.

[3] If E is constant during follow-up, this can be assessed by the unconditional association between Z and E in the data.

[4] If there is no selection bias (for E and Z) in the control population, the association in the source population can be estimated based on the controls alone.

Confounding — Real Example

Data: multicenter clinical trial on the efficacy of tolbutamide in preventing complications of diabetes,[5]

* Y = survival during follow-up period,
* E = treatment (tolbutamide/placebo),
* Z = age (dichotomized at 55 years).

Survival/ Treatment	All ages		Age < 55		Age ≥ 55	
	tolbut	placebo	tolbut	placebo	tolbut	placebo
Dead	30	21	8	5	22	16
Alive	174	184	98	115	76	69
risk (%)	14.7	10.2	7.5	4.2	22.4	18.8
RD (%)	4.5		3.3		3.6	
RR	1.44		1.81		1.19	
OR	1.51		1.88		1.25	

- could age be a confounder? — yes, because:
  * definitely a risk factor for mortality (within both E groups),
  * seems to be associated with treatment, despite the randomization (!),
  * not a result of neither E nor Y,
- does age have a confounding effect? — not so clear:
  * RR and OR for E not the same within age groups and totally, but total effect could be a "fair average".

[5] Data from University Group Diabetes Program, reproduced in Rothman *et al* (2008), *Modern Epidemiology*, 3rd ed., Table 15.1.

Counterfactual Definition of Confounding

Notation and setup: (following Greenland)

- aim: compare *treatment* x_1 with *reference* x_0 in *target population* A, in terms of population parameter μ,
- assume: actual effect under x_1 is μ_{A_1}, hypothetical effect under x_0 would be μ_{A_0},
- counterfactual effect: $\mu_{A_1} - \mu_{A_0}$, or μ_{A_1}/μ_{A_0} (for risks),
- assume: *reference population* B (as similar to A as possible) to which x_0 is applied, with actual effect μ_{B_0}

$\Rightarrow$ confounding exists whenever $\mu_{A_0} \neq \mu_{B_0}$, leading to biases in the (real) measure $\mu_{A_1} - \mu_{B_0}$ (or μ_{A_1}/μ_{B_0}).

Consequences:

- confounding depends on the outcome parameter (μ),[6]
- confounding depends on the target population of inference (A),
- although confounding is not defined for specific variables (Z), some variables must be distributed differently in A and B,
  - a confounder is a variable whose different distribution in A and B is "responsible" for the confounding.

[6] For example, choice of a 5-year or 10-year survival period; "we should generally expect no confounding for 200-year survival"... (Greenland).

Working Definition of Confounding

Problem: determining confounding from counterfactual reasoning or necessary conditions only is unmanageable,

- the required information is not available,
- a large pool of confounders may be anticipated, leading to overly complex analytical situations.

Pragmatic solution: define confounding of Z for the relation $E \rightarrow Y$ as present when conditions (i)–(ii) are met:

(i) Z meets the (3) necessary conditions to be a confounder,

(ii) the difference between a crude ("total") measure of association/effect and a confounding-adjusted measure of association/effect (see L10a–14) is "substantial", i.e.[7]

  * a bias above 20–30% (arbitrary cut-off set in VER) measured relative to the crude estimate.

Examples (revisited):

- constructed BRD data: unadjusted OR=3.32, adjusted OR=2.03, bias $= (3.32 - 2.03)/3.32 = 39\%$
  $\Rightarrow$ substantial confounding by BRSV,
- UGDP data: unadjusted RR=1.44, adjusted RR=1.33, bias $= (1.44 - 1.33)/1.44 = 8\%$
  $\Rightarrow$ no substantial confounding by age.

[7] Rothman *al* (2008), p. 262, note that usually 50% would be considered substantial, and 5% would not...

Non-Collapsibility of Odds-Ratios

Collapsibility in a 3-way (e.g., $E \times Y \times Z$) contingency table (cross-classification table of counts):

- a measure of association (ψ) of/between E and Y is strictly collapsible across Z if
  - ψ is constant across the conditional tables given Z (with Z at fixed values), *and*
  - ψ has the same value in the marginal $E \times Y$ table,
- a measure of association (ψ) of E and Y is collapsible if
  - the marginal ψ equals a (suitably chosen) weighted average of the ψ-values in the conditional tables (a confounding-adjusted ψ, see L10a–14),
- note: collapsibility depends on the chosen ψ.

Non-collapsibility (NC):

- also referred to as Simpson's or Yule-Berkson's paradox,
- what we observed in the previous examples! (L10a–4,6)
- is not the same as confounding![8]
  - NC possible without confounding: for $\psi = \text{OR}$ only (Example 13.6 of VER; not core material),
  - NC does not always follow from confounding: $\psi = \text{OR}$.

[8] When focusing on population effects rather than observed effects, the difference disappears (Greenland); as it does if the frequency outcome is low (VER).

Control of Confounding — Overview

3 major approaches to prevent confounding:

- exclusion/restriction: restrict the study population to one level of the potential confounder Z; for example,
  - ∗ select gender, breed, age (group),
  - ∗ reduce/prevent confounding by herd by selecting herds with similar production characteristics,

  note potential drawback: reduced scope of the study,
- matching: another approach based on study design, discussed in Lecture 11a,
- analytic control: procedures to account for any confounding effects in the data analysis, given a dataset where confounding is present; main coverage:[9]

  (VHM 811): stratification by Mantel-Haenszel procedures (L10a–14), whereby the strata are taken as the levels of the confounder,

  (VHM 812): multivariable modelling: include confounders among predictors in actual model.

Propensity scores (abbreviated as PS in VER):

= cond. prob. of being treated/exposed given covariates,

– used for: matching, stratification, multivar. modelling.

[9] Other approaches include: standardised risks/rates (13.7.1), marginal structural models (13.7.2), instrumental variables (13.9).

Graphical Analysis for Confounding

Motivation:
for complex causal structures, it can be difficult to assess which variables are potential confounders, and how analytic control for one confounder affects the situation for others.

Proposed graphical procedure from a causal diagram:[10]

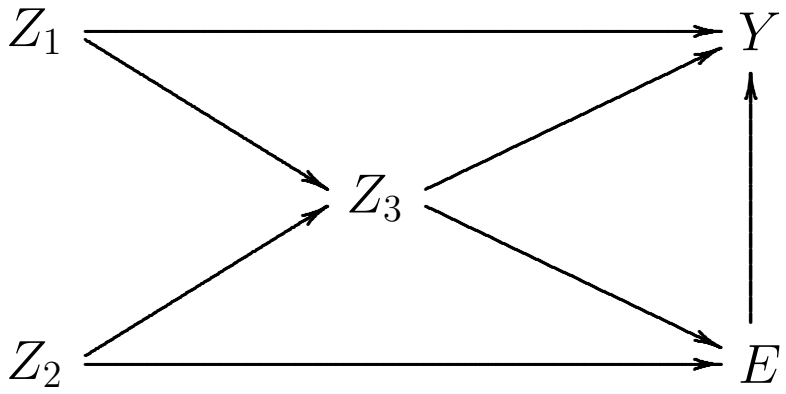
(i) delete all arrows away from E,

(ii) potential confounding $\sim$ unblocked paths $E \to Y$
(a blocked path has arrows "pointing at each other"),

plus a rule on controlling for confounding:

(iii) control of a variable may create an association between any variables with arrows into the corresponding node
$\Rightarrow$ reassessment of the causal diagram is needed!

Example for (iii):[9]



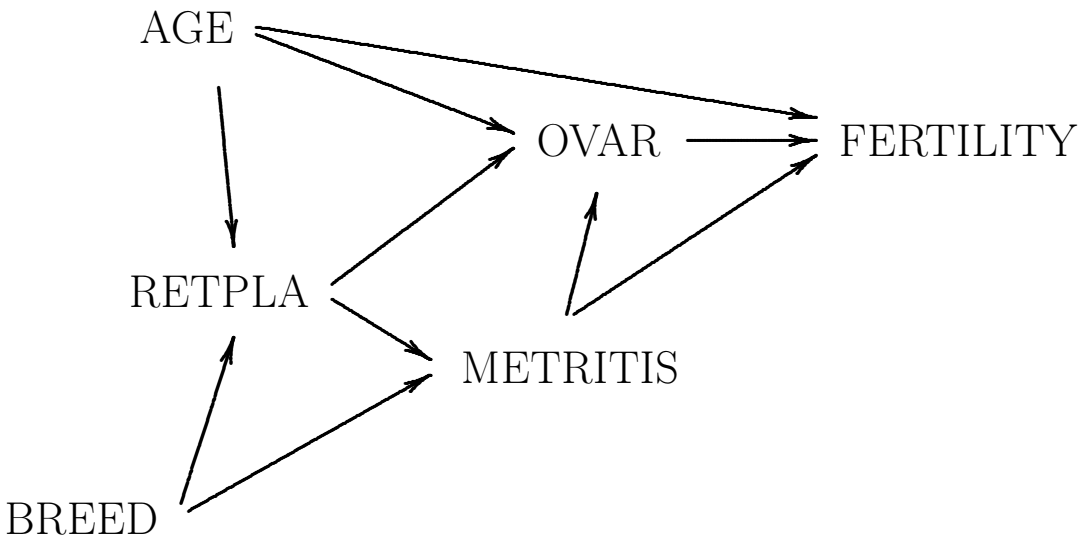
Control of Z_3
is insufficient!

[10] Details can be found in Greenland S, Pearl J, Robins JM (1999), Causal diagrams for epidemiologic research, *Epidemiology* **10**, 37–48.

Example of Multifactorial Graph

Infertility in dairy cows (Examples 1.4 & 13.5)

* outcome Y: measure of fertility (e.g., calving to conception interval),
* exposure E of interest: metritis,



Graphical analysis for relation `METRITIS` → `FERTIITY`:

(i) remove two arrows away from `METRITIS`,

(ii) unblocked paths from `METRITIS` to `FERTILITY` exist via `OVAR` and `AGE`,

(iii) both paths go through `RETPLA` ⇒ should be controlled,

(iv) control of `RETPLA` creates extra association between `AGE` and `BREED`, and new unblocked path ⇒ one of these should be controlled as well.

Illustration: Association Created by Control

Context: Rule (iii) for graphical assessment of confounding (L10a–11) states that control of one variable may create an association between its "ancestors" (variables affecting it).

Numerical illustration (artificial data)[11] of how stratification may create association; refer to figure on L10a–11 for notation:

	$Z_1 = 1$		$Z_1 = 2$	
	$Z_2 = 1$	$Z_2 = 2$	$Z_2 = 1$	$Z_2 = 2$
$Z_3 = 1$	800	600	400	200
$Z_3 = 2$	200	400	600	800
total	1000	1000	1000	1000

- Z_1 and Z_2 have no association (total row),
- within both strata of Z_3, the odds-ratio for Z_1 and Z_2 equals $800 \cdot 200/(400 \cdot 600) = 0.67$ (so association exists),
- Z_1 and Z_3 are associated:
  $\text{OR} = (800 + 600) \cdot (600 + 800)/((200 + 400) \cdot (400 + 200)) = 5.44$,
- Z_2 and Z_3 are associated:
  $\text{OR} = (800 + 400) \cdot (400 + 800)/((200 + 600) \cdot (600 + 200)) = 2.25$,
- "When we stratify on Z_3, the association of Z_1 and Z_2 within at least one stratum of Z_3 will almost certainly differ from the crude $Z_1 - Z_2$ association".[11]

[11] Example is in Table 1 of Greenland, Pearl & Robins (1999).

Mantel-Haenszel Procedures

- name[12] of a range of statistical procedures to combine information from multiple strata into estimates and tests,
- variants exist for multiple data types, study designs and effect measures; consider here the odds-ratio (OR),
- main advantage versus alternative weighting schemes (including multivariable methods) combining estimates across multiple strata, is a robustness in sparse data.[13]

Notation and formulae for M-H estimate and test for OR:[14]

Stratum j	$E+$	$E-$	Total
$Y = 1$ (case)	a_{1j}	a_{0j}	m_{1j}
$Y = 0$ (control)	b_{1j}	b_{0j}	m_{0j}
Total	n_{1j}	n_{0j}	n_j

$\mathrm{OR}_j = [a_{1j}b_{0j}]/[a_{0j}b_{1j}]$
$E_j = m_{1j}n_{1j}/n_j$
$V_j = [m_{1j}m_{0j}n_{1j}n_{0j}]/[n_j^2(n_j - 1)]$

E_j = expected no. cases; V_j = variance of expected no. cases

- M-H odds-ratio: $\mathrm{OR}_{\mathrm{MH}} = \dfrac{\sum a_{1j}b_{0j}/n_j}{\sum a_{0j}b_{1j}/n_j}$,
- M-H test of OR = 1: $X^2_{\mathrm{MH}} = [\sum a_{1j} - \sum E_j]^2 / \sum V_j$, $\sim$ (approx.) χ^2-distribution with 1 df under H_0.

[12] From a seminal paper: Mantel N, Haenszel WH (1959), Statistical aspects of the analysis of data from retrospective studies of disease, *J Natl Cancer Inst* **22**, 719–48.

[13] The statistical inference typically requires tables for each stratum to be not too small, but the requirements for M-H statistics are primarily for summarized counts across all strata; detailed coverage in (e.g.) Rothman *et al* (2008), Chapter 15; Kleinbaum, Kupper & Morgenstern (1982), *Epidemiologic Research: Principles and Quantitative Methods*, Chapter 17.

[14] Formulae valid for cross-sectional, cohort, and case-control designs.

Dataset FEEDLOT — Confounding by Province

- case-control data pooled from multiple studies,
- 588 animal obs. based on first 28 days at feedlot.

Variable	Description	Values
brd	clinical BRD status	0/1 (control/case)
phcysc	seroconv. to Mh during study	0/1
ibrsc	seroconv. to IBR during study	0/1
brsvsc	seroconv. to BRSV during study	0/1
province	province of feedlot	1/2 (Alberta/Ontario)

M-H analysis for confounding by province (Example 13.8):

BRD status/ Exposure	Combined		Alberta		Ontario	
	Mh +	Mh −	Mh +	Mh −	Mh +	Mh −
case	167	30	84	21	83	9
control	300	91	80	55	220	36
OR (95% CI)	1.69 (1.05,2.76)		2.75 (1.47,5.22)		1.51 (0.68,3.72)	
OR_{MH} (95% CI)			2.19 (1.37,3.51)			
X^2_{MH} (P-value)			11.2 (.0008)			

- confounding by province?: $(2.19-1.69)/1.69 = 30\%$
  $\Rightarrow$ moderate/substantial (potential) confounding,
- significance for Mh effect?: yes, strongly so ($P < 0.001$),

$\Rightarrow$ control for province recommended, and OR_{MH} is our preferred estimate.

Feedlot Data — M-H Analysis for IBR

Demonstration of M-H calculations (Example 13.7):[15]

Stratified tables:

BRD status/ Exposure	IBR +			IBR –		
	Mh +	Mh –	total	Mh +	Mh –	total
case	83	18	83	84	12	96
control	85	48	220	215	43	256
total	168	66	234	299	55	354

Strata-specific calculations:

Stratum j	n_j	$a_{1j}b_{0j}/n_j$	$a_{0j}b_{1j}/n_j$	OR_j	a_{1j}	E_j	V_j
IBR=1	234	17.03	6.54	2.60	83	72.51	11.67
IBR=0	354	10.20	7.29	1.40	84	81.08	9.21
sum	–	27.23	13.83	–	167	153.60	20.88

- $\text{OR}_{\text{MH}} = 27.23/13.83 = 1.97$,
- confounding by IBR?: $(1.97-1.69)/1.69 = 17\%$
  $\Rightarrow$ minor (potential) confounding,
- $X^2_{\text{MH}} = (167–153.60)^2/20.88 = 8.60$, strongly significant in $\chi^2(1)$ with $P = 0.0034$.

$\Rightarrow$ confounding by IBR perhaps ignorable and no major discrepancy between crude and adjusted OR's (all of which should be reassessed while accounting for province).

[15] I can hardly think of any reason to not use software in a practical situation. Also, after confounding by province has been demonstrated, it would be inconsistent to revert to analyses ignoring province.

Stata Commands for M-H Analysis

- uses same commands as previously for different designs (`cs` and `cc`; note that the interactive forms `csi` and `cci` don't allow for a third variable),
- M-H analysis stratified across variable `z` only requires the added option `by(z)`, for example for Example 13.8:

  ```
  cc brd mh, by(province)
  ```

  ,
- stratified analysis possible for both RR and OR (but apparently not for RD with M-H method),
- other weighting schemes than M-H weights are possible, in particular internal and external standardization (options `estandard` and `istandard`),
- conditional tables for each level of Z are most easily produced using restrictions in `if` clauses, for example for Example 13.8:

  ```
  cc brd mh if province==1
  ```

Data entry and data formats:

- for entry of a table of counts, one may type values into the Stata editor, or construct the data in Excel (or similar) and import as a comma-separated file (`insheet` command);
  sample data structure (Example 13.1):

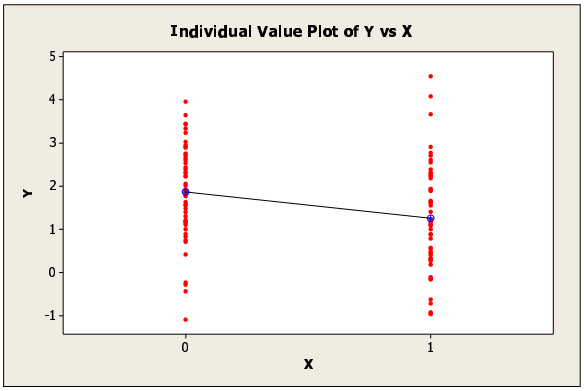
```
brd   mh   brsv    n
 1    1      1    220
 1    0      1     10
 0    1      1   5280
...
```

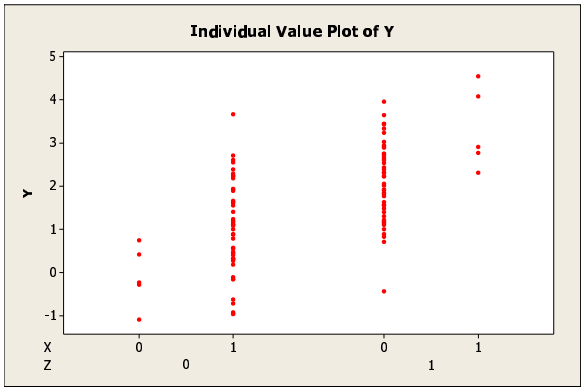
- for frequency-tabulated data (e.g. the sample data structure above), add `[fw=n]` to `cc` and `cs` commands.

Confounding — Other Situations

Confounding is for binary outcomes and predictors only? Absolutely not! — some examples:

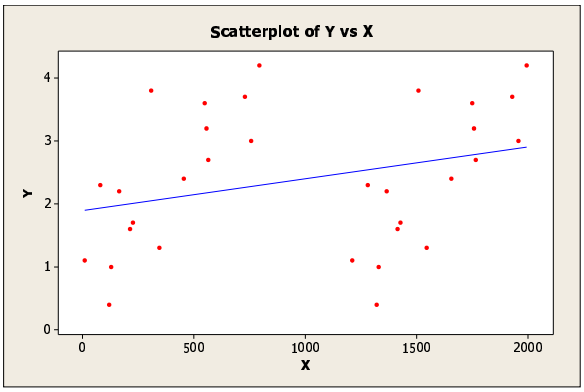
- categorical confounder: no problem, just split into more strata (Mantel-Haenszel procedure still applies),
- continuous outcome: two-sample comparison confounded by a third variable,

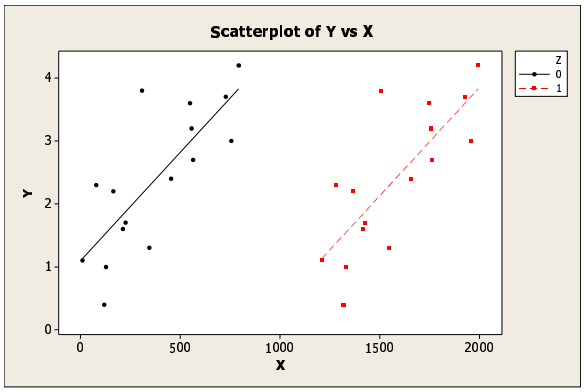




overall, largest mean for $X = 0$, but opposite order within levels of Z (all associations significant),

- continuous outcome, continuous predictor: comparison of linear regressions,





steeper slope for regression within Z levels — confounding?

Stata Do-file

```
* Example 13.1
insheet using "eg13_01.csv"
cs brd mh [fw=n], or
cs brd mh [fw=n] if brsv==1, or
cs brd mh [fw=n] if brsv==0, or
cs brd mh [fw=n], or by(brsv)

* Diabetes trial
use http://www.stata-press.com/data/r10/ugdp, clear
cs case exposed [fw=pop], or
cs case exposed [fw=pop] if age==0, or
cs case exposed [fw=pop] if age==1, or
cs case exposed [fw=pop], by(age)

* feedlot data
use feedlot.dta, clear
rename phcysc mh
tab brd mh
* Example 13.8
cc brd mh
cc brd mh if province==1
cc brd mh if province==2
cc brd mh, by(province)
* Example 13.7
cc brd mh if ibrsc==1
cc brd mh if ibrsc==0
cc brd mh, by(ibrsc)
```