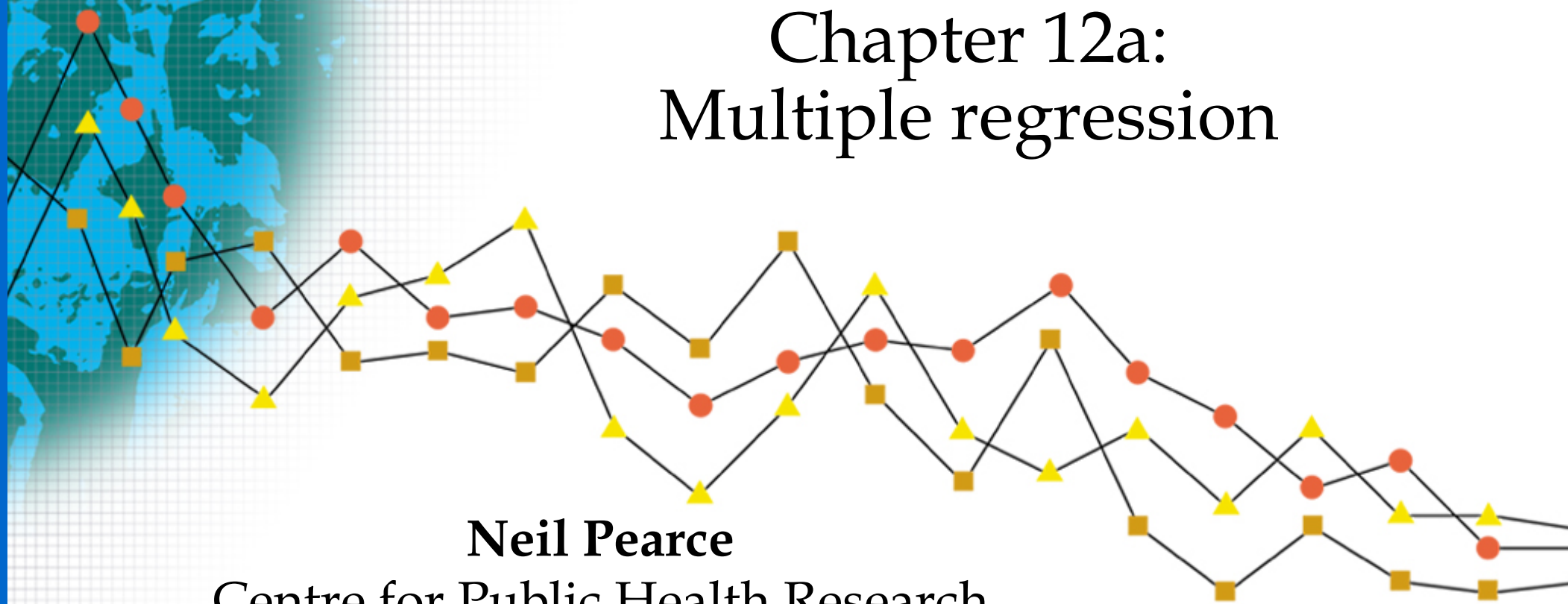
A short introduction to epidemiology

Chapter 12a: Multiple regression

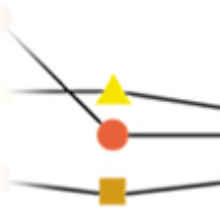


Neil Pearce
Centre for Public Health Research
Massey University
Wellington, New Zealand

Chapter 9 (additional material)

Multiple regression

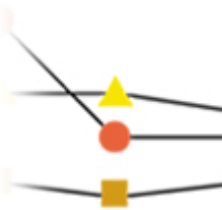
- This presentation includes additional material on data analysis using multiple regression



Chapter 9 (additional material)

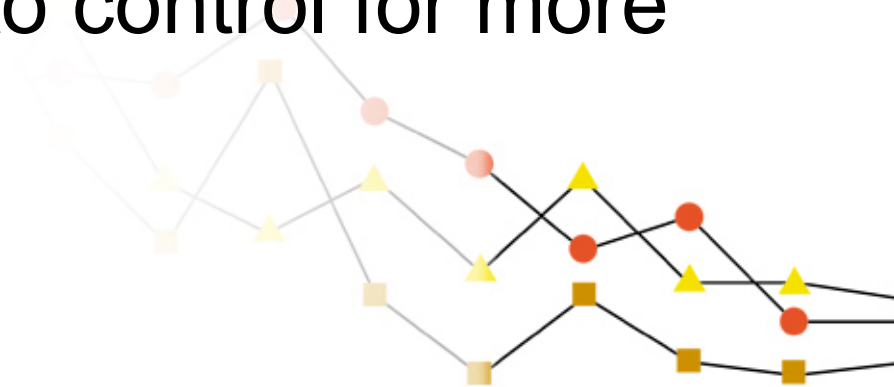
Multiple regression

- Why use multiple regression?
- The basic regression model
- Interaction
- Model selection
- Regression diagnostics
- Approaches to regression



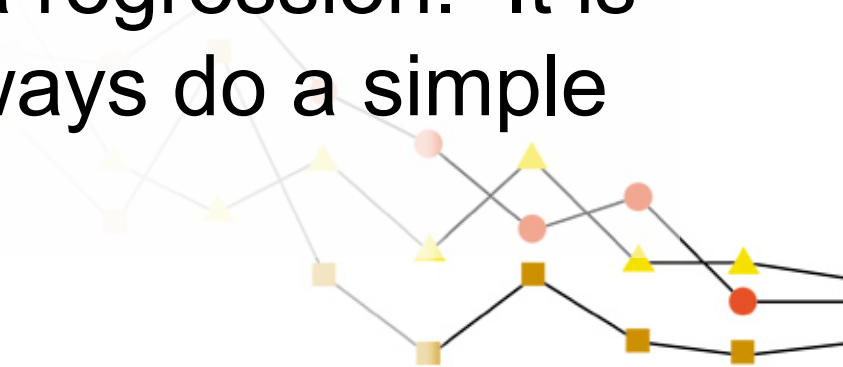
Why Use Multiple Regression?

- Regression models produce estimates which are both statistically optimal and mutually standardized
- Stratification (or adjustment through stratification) will have problems with small numbers if it is necessary to control for more than 2 or 3 confounders



Some Reasons for Caution With Regression

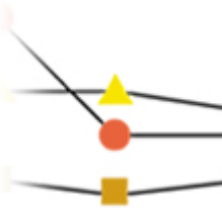
- The gain in statistical efficiency occurs because the model makes certain assumptions about the structure of the data. These assumptions may be wrong
- You have less control and understanding of the analysis when you use a regression. It is easy to make mistakes. Always do a simple stratified analysis first



Chapter 9 (additional material)

Multiple regression

- Why use multiple regression?
- The basic regression model
- Interaction
- Model selection
- Regression diagnostics
- Approaches to regression



Regression

	Exposed	Non-exposed
Deaths	18,000	9,500
Person-years	900,000	950,000

$$I_1 = \frac{18{,}000}{900{,}000} = 0.02 \text{ deaths per person}$$

$$I_0 = \frac{9{,}500}{950{,}000} = 0.01 \text{ deaths per person - year}$$

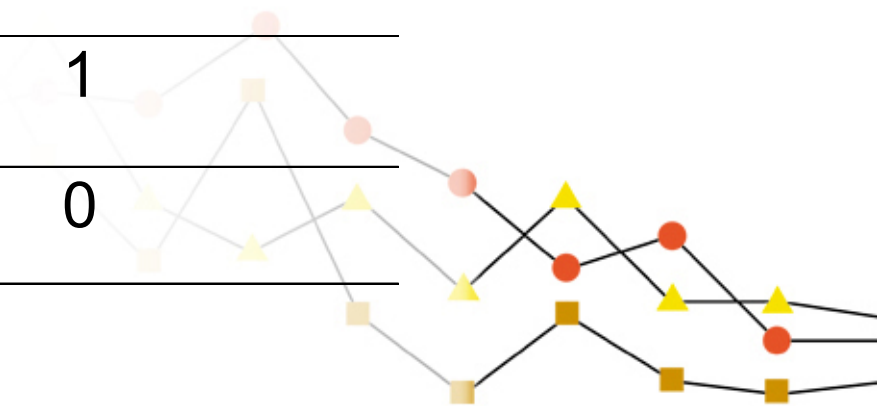
$$RR = \frac{I_1}{I_0} = \frac{0.02}{0.01} = 2.00$$

We can achieve the same result by using a regression model. We define a dichotomous exposure variable (X_1) as:

Exposed: $X_1 = 1$

Non-exposed: $X_1 = 0$

Rate (I)	Exposure (X_1)
0.02	1
0.01	0

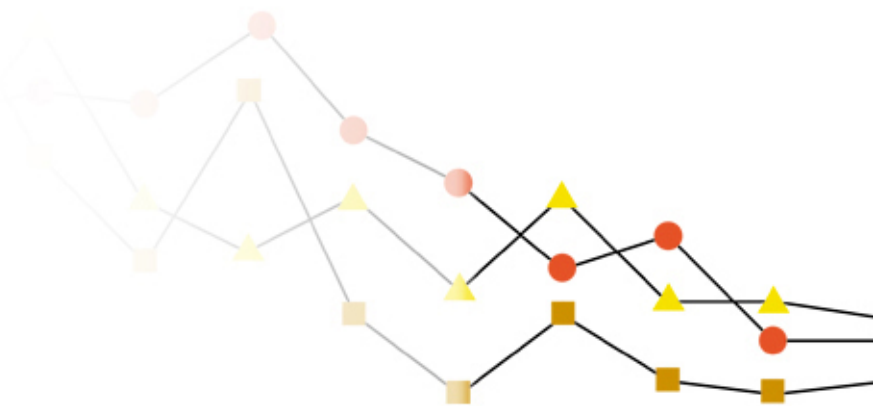


We want to model the rate (I) as a function of exposure (X_1).

One possibility is:

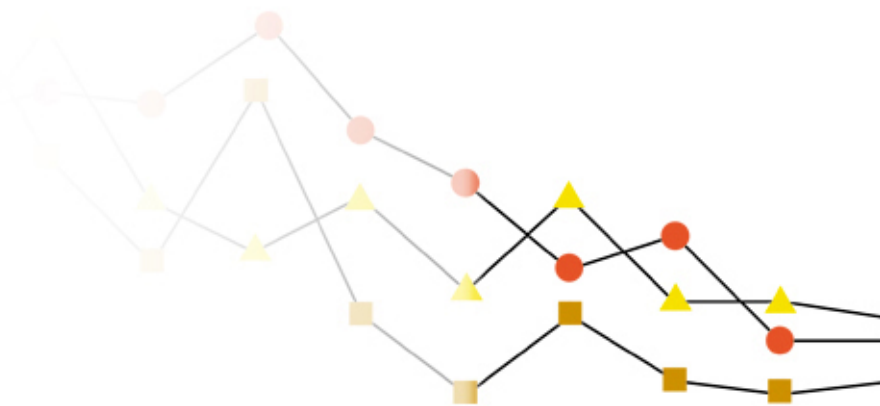
$$I = b_0 + b_1 X_1 (+E)$$

but this is less convenient statistically

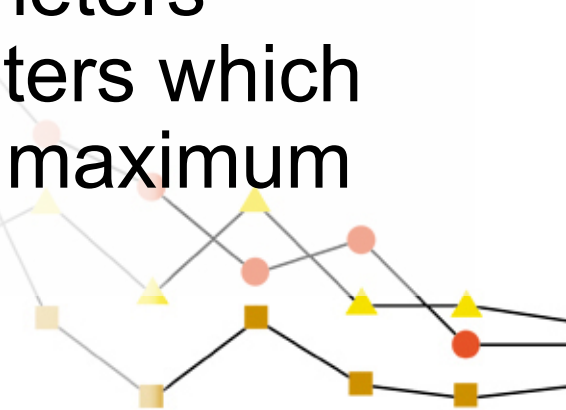


It is more convenient to fit the model:

$$\ln(I) = b_0 + b_1 X_1 (+E)$$



We could fit the model using simple linear regression (least squares). However, the least-squares approach does not handle Poisson or dichotomous outcome variables well, as they are not normally distributed. Instead, the model parameters are estimated by the method of *maximum likelihood*. This is based on the *likelihood function* which represents the probability of observing the actual data as a function of the unknown parameters (b_0,b_1,b_2, …). The values of the parameters which maximize the likelihood function are the maximum likelihood estimates of the parameters



Suppose we fit this model and obtain estimates for b_0+b_1

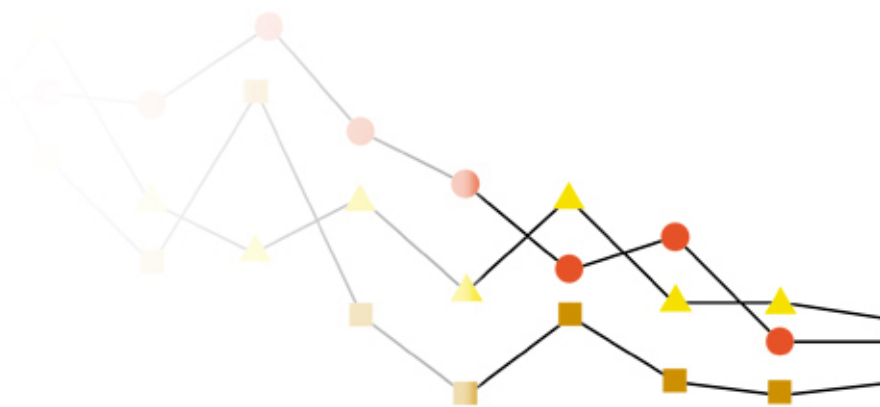
$$\text{Exposed}: \ln(I_1) = b_0 + b_1 \qquad (X_1 = 1)$$

$$\text{Non}-\text{exposed}: \ln(I_0) = b_0 \qquad (X_1 = 0)$$

$$\ln(I_1) - \ln(I_0) = (b_0 + b_1) - b_0 = b_1$$

$$\ln(I_1 / I_0) = b_1$$

$$\ln(RR) = b_1 \Rightarrow RR = e^{b_1}$$



The 95% CI for ln(RR) is:

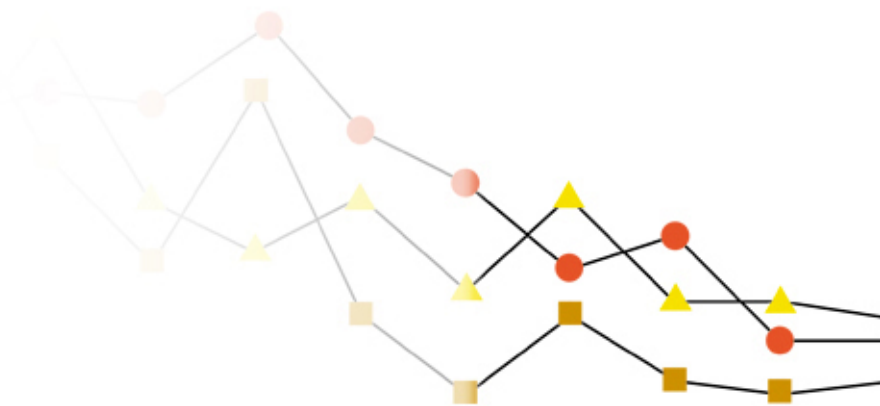
$$\ln(RR) \pm 1.96 SE[\ln(RR] = b_1 \pm 1.96.SE(b_1)$$

e.g. $b_1 = 0.693$, $SE(b_1) = 0.124$

$$RR = e^{b_1} = 2.00$$

$$95\%\ \text{lower limit} = e^{0.693 - 1.96 x 0.124} = 1.63$$

$$95\%\ \text{upper limit} = e^{0.693 + 1.96 x 0.124} = 2.45$$

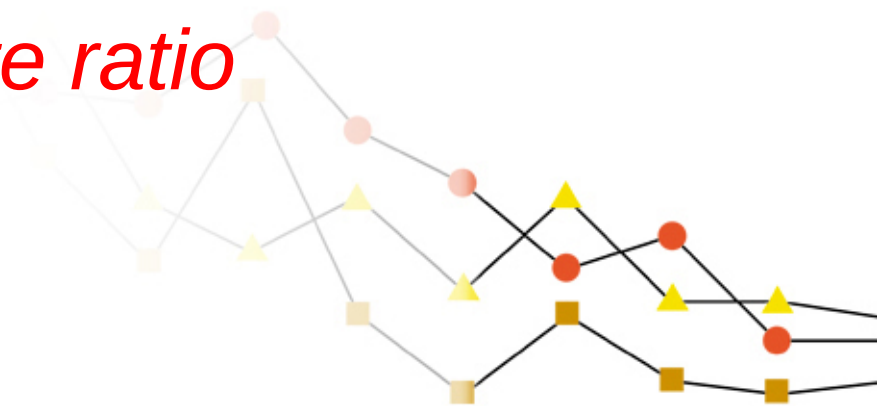


This general approach can be used in a variety of situations.

For *cohort studies* we fit the model

$$\ln(I) = b_0 + b_1 \, \mathrm{X}_2$$

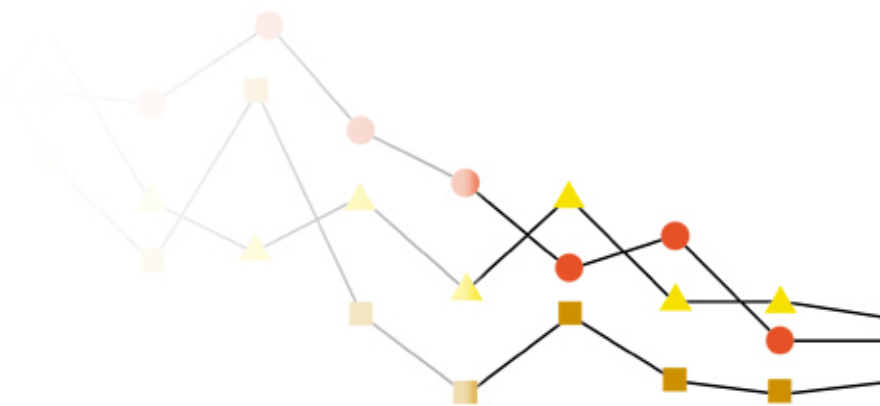
This is Poisson data, and we use *Poisson regression* to estimate the *rate ratio*



For *case-control studies* we fit the model

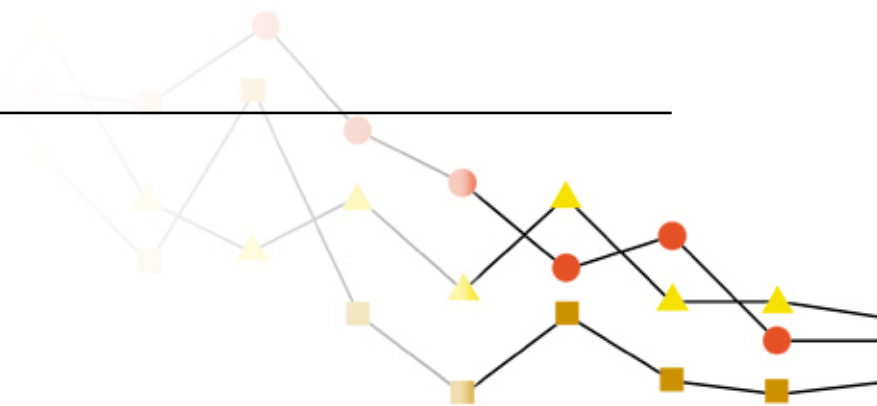
$$\ln(P/(1-P)) = b_0 + b_1 \, \mathrm{X}_1 + \ldots$$

This is logit (*binomial*) data and we use *logistic regression* to estimate the *odds ratio*



We can use the same approach to control for potential confounding variables:

	Age <50		Age≥50	
	E	$\bar{E}$	E	$\bar{E}$
Deaths	6,000	3,000	12,000	6,500
Person-years	400,000	450,000	500,000	500,000



We define X_1=1 (exposed)

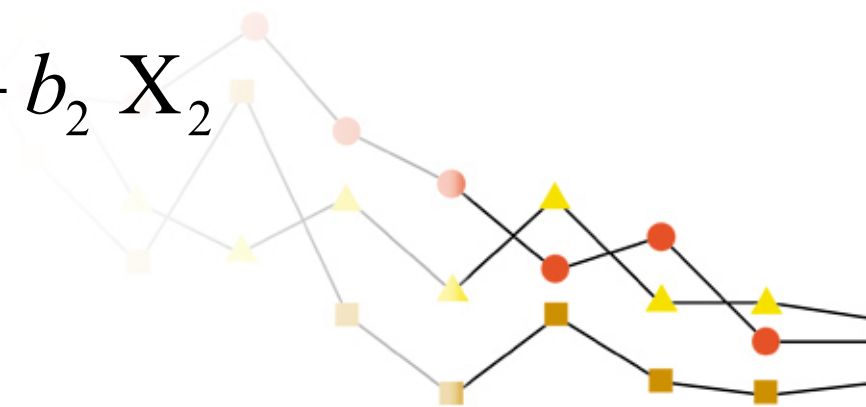
=0 (non-exposed)

X_2=1 (Age≥50)

=0 (Age <50)

We then run the model

$$\ln(I_1) = b_0 + b_1 X_1 + b_2 X_2$$



Then in the exposed group:

$$\ln(I_1) = b_0 + b_1 + b_2\ \mathrm{X}_2 \qquad (\mathrm{X}_1 = 1)$$

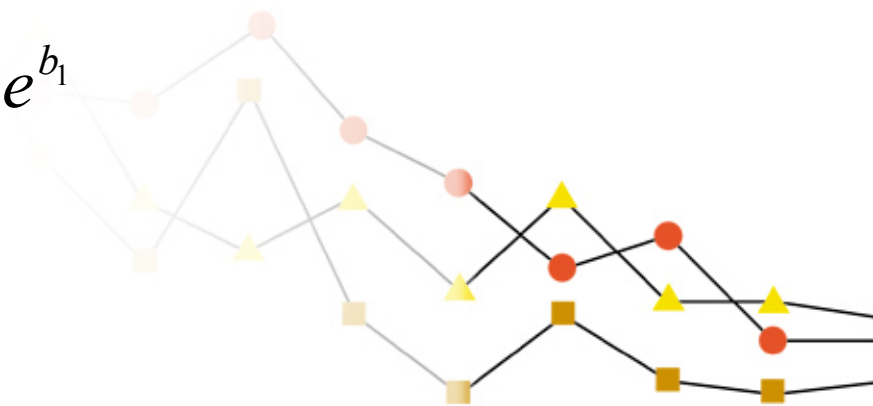
And in the non-exposed group:

$$\ln(I_0) = b_0 + b_2\ \mathrm{X}_2 \qquad (\mathrm{X}_1 = 0)$$

$$\Rightarrow \ln(\frac{I_1}{I_0}) = b_0 + b_1 + b_2\ \mathrm{X}_2 \text{ - } (b_0 + b_2\ \mathrm{X}_2)$$

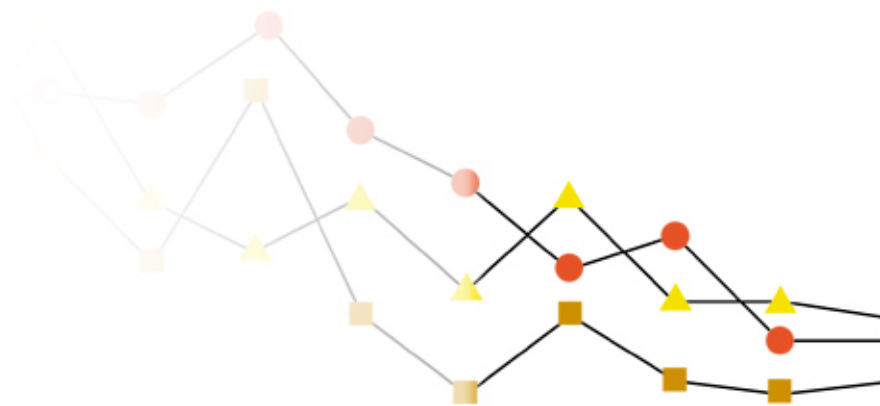
$$\Rightarrow \ln(RR) = b_1 \qquad RR = e^{b_1}$$

and we proceed as before



Multiple Levels

- We can also represent multiple categories of exposure (or a confounder): suppose we have four levels of exposure: none, low, medium, high
- We need *three* variables to represent *four* levels of exposure:



We fit the model:

$$\ln(I) = b_0 + b_1 \, \mathrm{X}_1 + b_2 \, \mathrm{X}_2 + b_3 \mathrm{X}_3 + \ldots$$

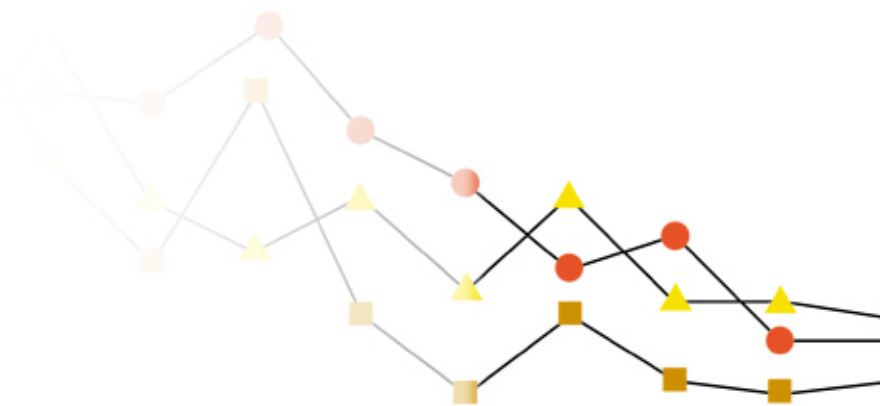
$$low \Leftrightarrow none : \ln(RR) = b_1$$

$$medium \Leftrightarrow none : \ln(RR) = b_2$$

$$high \Leftrightarrow none : \ln(RR) = b_3$$

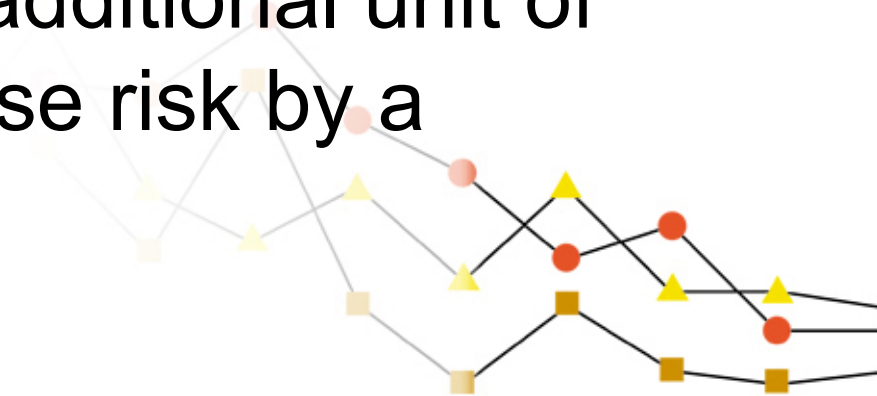
We can thus estimate the risk for each level relative to the lowest level of exposure.

We can control for confounding in a similar way, eg by defining *five* variables to represent *six* age-groups

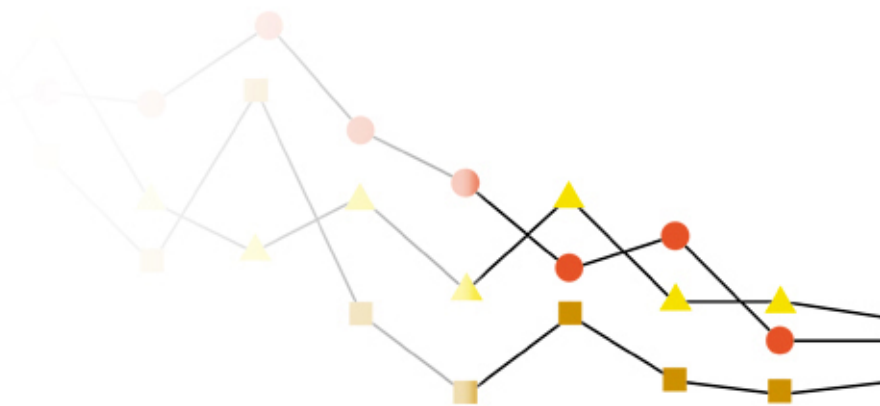


Rather than categorizing exposures it is possible to use each inidividual's exact exposure and to represent exposure with a single continuous variable.

However, the use of a continuous variable assumes that exposure is exponentially related to disease risk, ie, that each additional unit of exposure multiplies the disease risk by a certain amount.



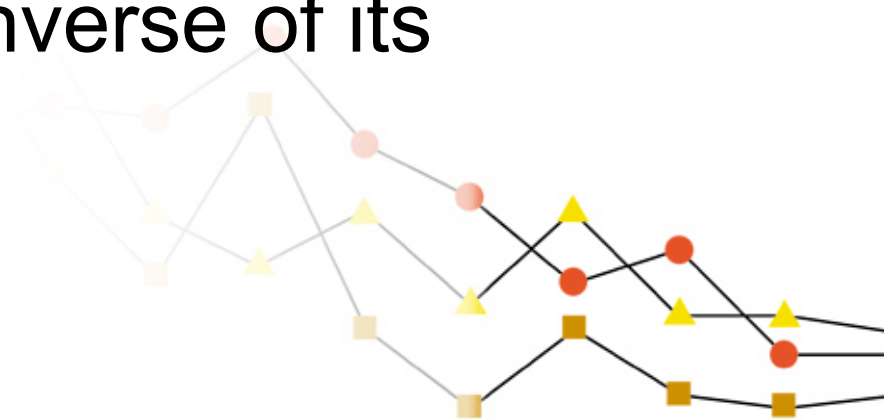
In other words, it assumes that the dose-response curve looks like this:

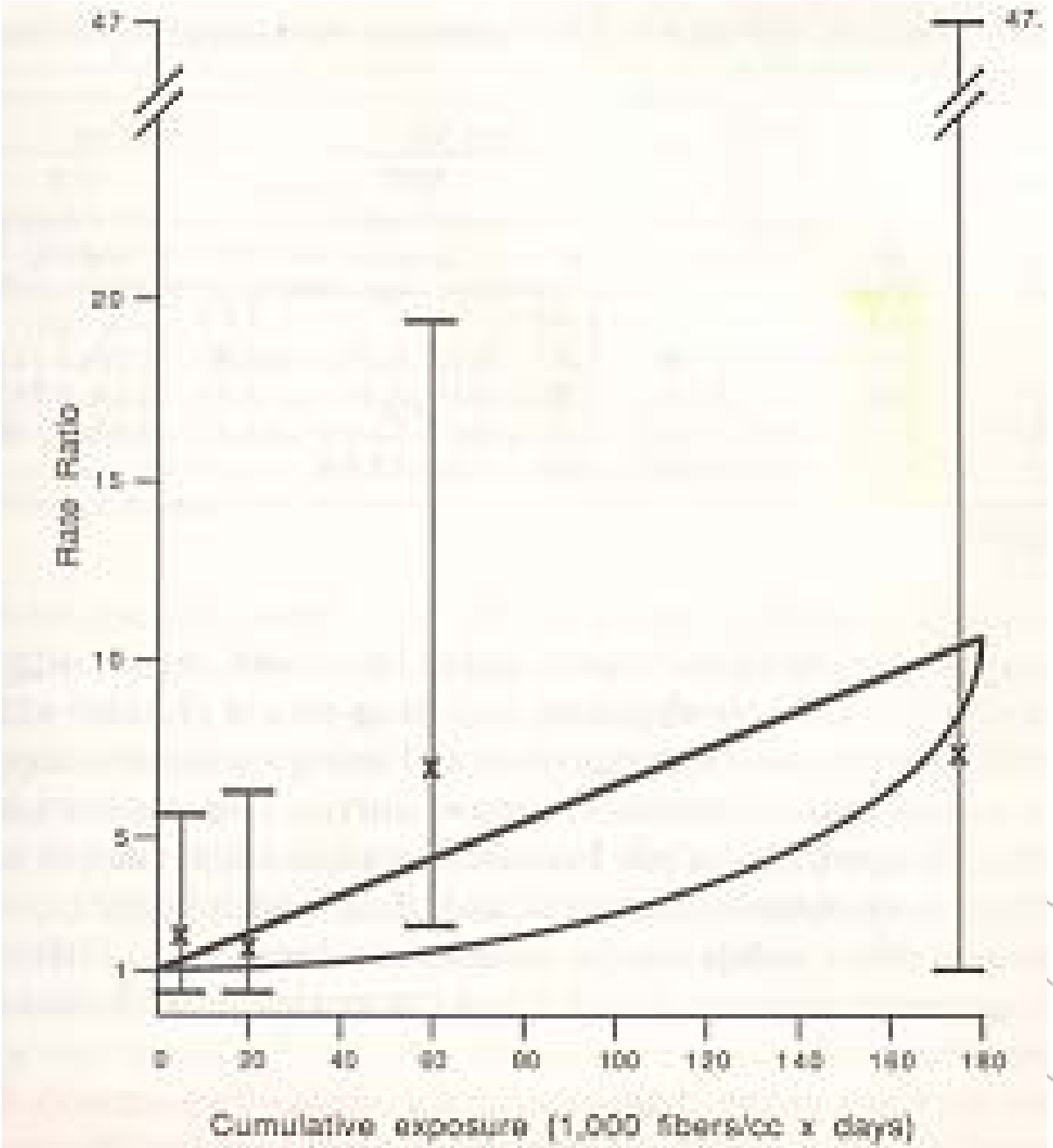


This assumption will not be optimal if the true dose-response curve is linear, or some other non-expondential shape.

There is little loss of statistical power providing it is possible to use at least 4 categories, and categorization is thus preferable as it provides for a greater understanding of the findings.

Appropriate methods do exist for modeling the close-response curve in an appropriate fashion once the appropriate shape of the curve has been determined. This generally involves taking the relative risk estimates of each of the individual exposure categories and performing an ordinary linear regression where each estimate is weighted by the inverse of its variance.

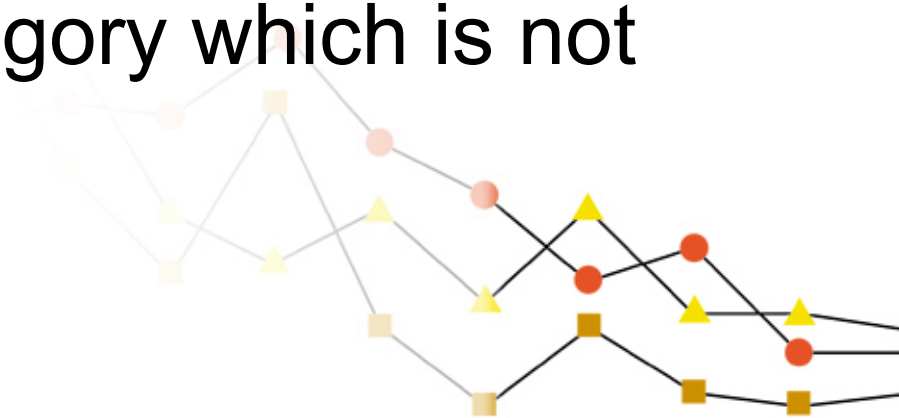



47
47
20
15
10
5
1
Rate Ratio
0
20
40
60
80
100
120
140
160
180
Cumulative exposure (1,000 fibers/cc x days)

Confounders

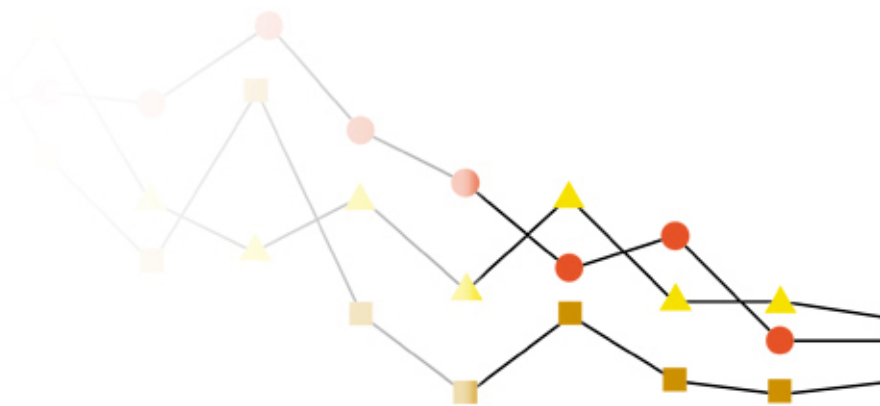
The same considerations aply to the definition of confounders.

For example, if there are 5 age-groups then we need 4 dummy variables [one of the age-groups, usually the youngest one, is taken as the baseline “reference” category which is not represented by a variable.]



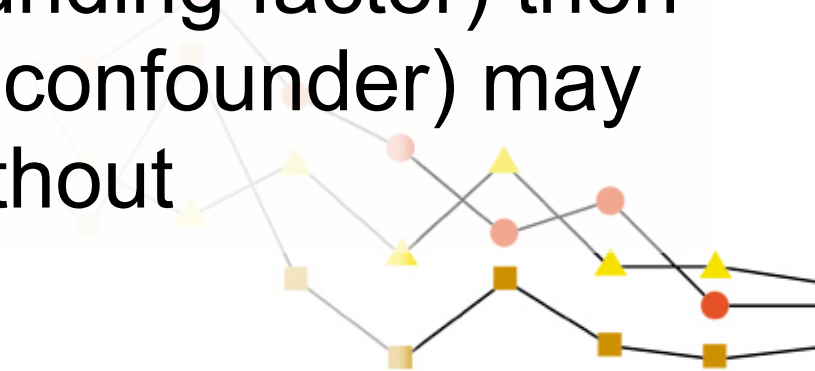
The model would then look like this:

$$\ln(I) = \underbrace{B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4}_{\text{Exposure variables}} + \underbrace{B_5X_5 + B_6X_6 + B_7X_7 + B_8X_8}_{\text{Age groups}}$$



Once again, it is preferable to use categorical rather than continuous variables to adjust for confounders. However, the issue is not so important, since the intention is simply to “adjust for the confounder rather than model its dose-response relationship.

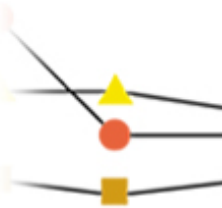
However, if our aim is simply to control confounding (rather than to estimate the dose-response pattern for the confounding factor) then an continuous variable (for the confounder) may be more statistically optimal without compromising validity



Chapter 9 (additional material)

Multiple regression

- Why use multiple regression?
- The basic regression model
- Interaction
- Model selection
- Regression diagnostics
- Approaches to regression



Interaction (Joint Effects)

Suppose that we wish to derive the following table:

Asbestos (X_1)	Smoking (X_2) Yes	Smoking (X_2) No
Yes	R_{11}	R_{01}
No	R_{10}	1.0

The usual model (without an interaction term) is:

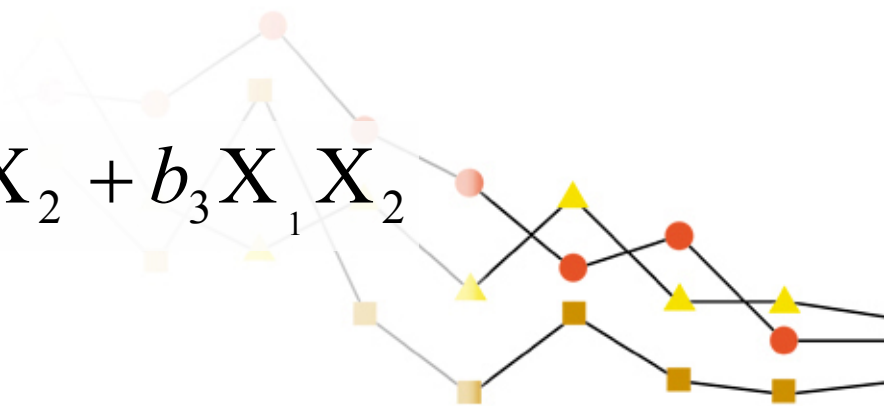
$$\ln(I) = b_0 + b_1 \mathrm{X}_1 + b_2 \mathrm{X}_2$$

$$\downarrow \qquad \downarrow$$

$$(\text{Asbestos})(\text{ Smoking})$$

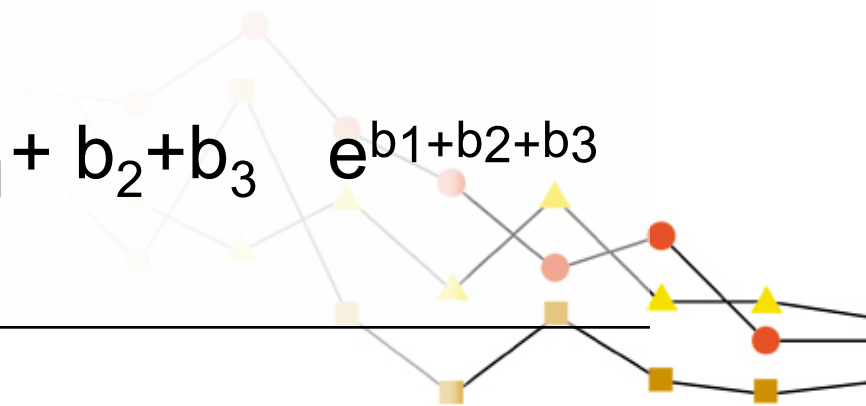
However, to get the above table, we need to fit the following model:

$$\ln(I) = b_0 + b_1 \mathrm{X}_1 + b_2 \mathrm{X}_2 + b_3 \mathrm{X}_1 \mathrm{X}_2$$



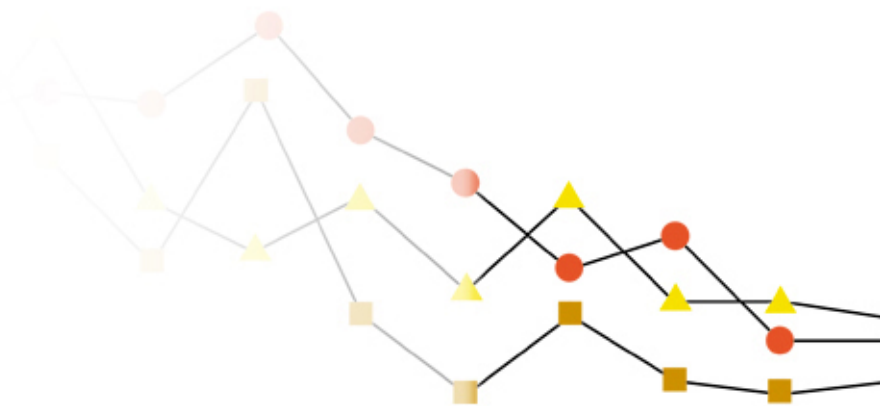
This can be used to derive the following:

	Asbestos	Smoking		
Group	X_1	X_2	Model	RR
Neither	0	0	b_0	1.0
Asbestos only	1	0	b_0+b_1	e^{b1}
Smoking only	0	1	b_0+b_2	e^{b2}
Both	1	1	$b_0+ b_1+ b_2+b_3$	$e^{b1+b2+b3}$

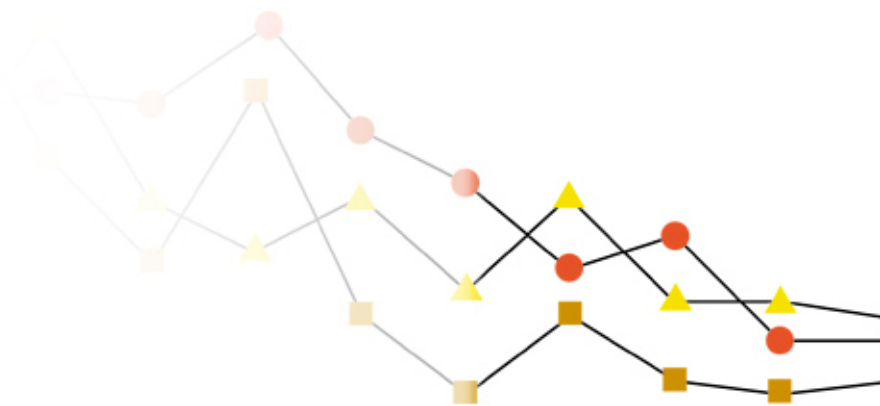


Thus, the joint effect is obtained by

$$e^{b_1+b_2+b_3} = e^{b_1} \cdot e^{b_2} \cdot e^{b_3}$$



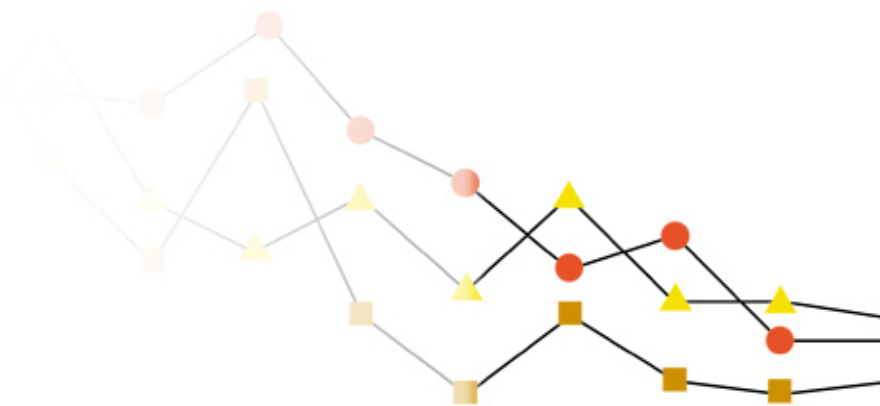
Note that if $b_3=0$ then the joint effect is just $e^{b1}.e^{b2}$. Thus, b_3 provides a test for interaction. However, it is important to emphasize that b_3 only provides a test for a departure from the mulitplicative assumptions of the model. It does not test for a departure from additivity.



Unfortunately, calculating the confidence interval for the joing effect is also complicated. We use:

$$\mathrm{Var}(b_1 + \mathrm{b}_2 + \mathrm{b}_3) = \mathrm{Var}(b_1) + \mathrm{Var}(\mathrm{b}_2) + Var(b_3) + 2[\mathrm{Cov}(\mathrm{X}_1, \mathrm{X}_2) + Cov(\mathrm{X}_1\mathrm{X}_3) + Cov(\mathrm{X}_2\mathrm{X}_3)]$$

$$and\ SE(b_1 + \mathrm{b}_2 + \mathrm{b}_3) = \sqrt{Var(b_1 + \mathrm{b}_2 + \mathrm{b}_3)}$$



There is a much easier way to get the same results. Just define three new variables as follows:

X_1 =1 if asbestos but not smoking

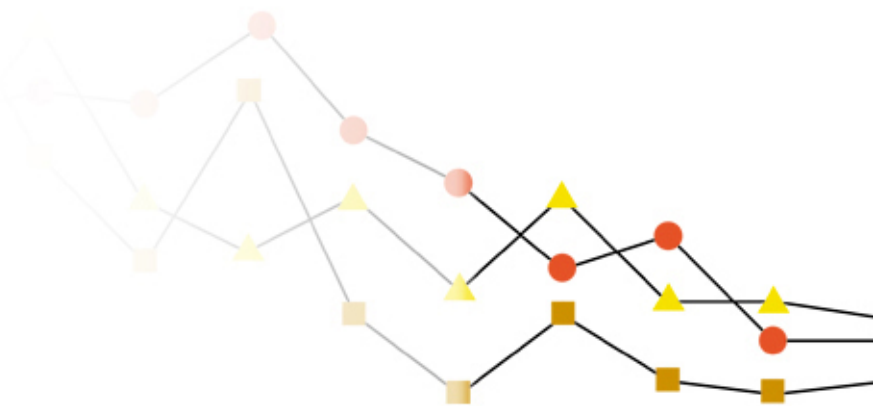
=0 otherwise

X_2=1 if smoking but not asbestos

=0 otherwise

X_3=1 if both

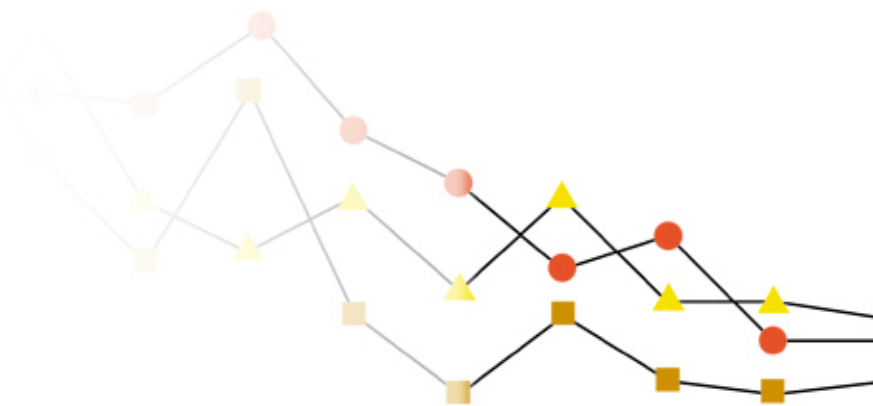
=0 otherwise



Then fit:

$$\ln(\mathrm{X}) = b_0 + b_1\ \mathrm{X}_1 + \mathrm{b}_2\mathrm{X}_2 + \mathrm{b}_3\ \mathrm{X}_3$$

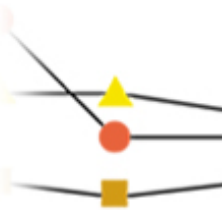
This will give us the separate and joint effects directly without any need to consider the Variance-covariance matrix.



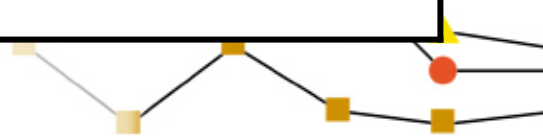
Chapter 9 (additional material)

Multiple regression

- Why use multiple regression?
- The basic regression model
- Interaction
- Model selection
- Regression diagnostics
- Approaches to regression



	Cohort study	Case control study
Numerator	Cases	Cases
Denominator	Person-Years	Controls
Effect estimate	Rate ratio	Odds ratio
Stratified analysis	SRR (or Mantel-Haenzsel)	Mantel-Haenszel
Modelling	Poisson regression	Logistic regression
Model	$\ln(I)=b_0+ b_1 + X_1+\ldots$	$\ln(P/(1-P))= b_0+ b_1 + X_1+\ldots$
Data structure	Poisson	Logit (binominal)
Programs	Stata or SAS	Stata or SAS



Use of Multiple regression

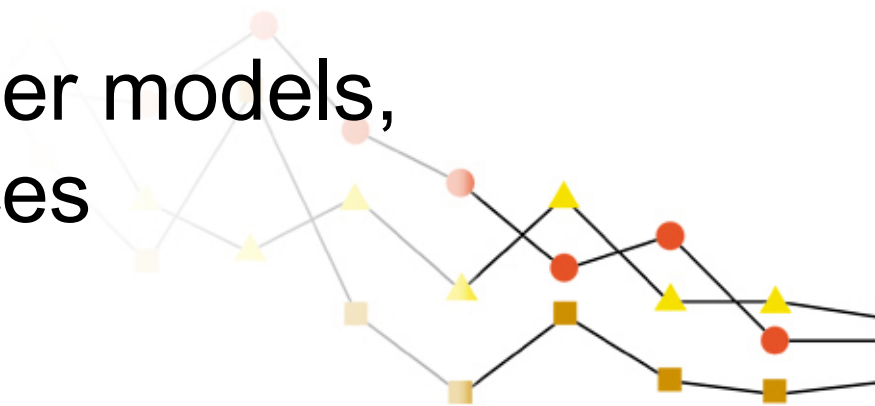
- Don’t use a regression model unless there is a good reason to do so
- The most common reason to use a model is because you need to simultaneously adjust for 4 or more confounders
- Most analyses can be handled with a simple stratified analysis and the Mantel-Haenszel summary odds ratio or rate ratio

Use the regression model which is appropriate for the data you have: don't make the data adapt to the model

Poisson regression is the appropriate model for cohort studies with incidence rates

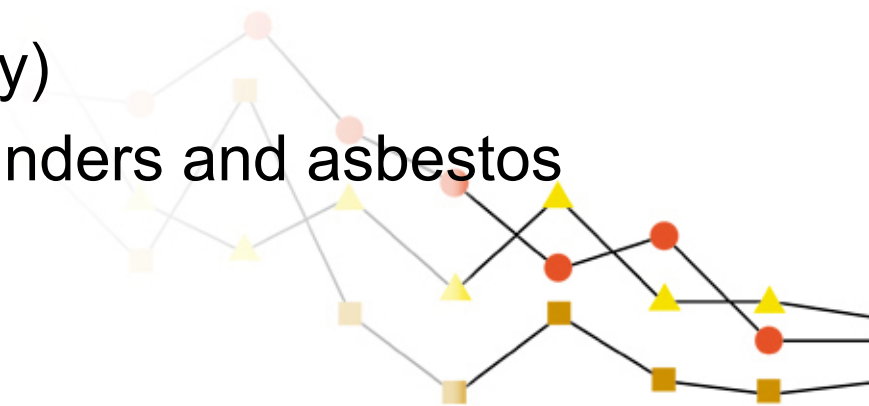
Logistic regression is the appropriate model for case-control data

There is no reason to use other models, except in special circumstances



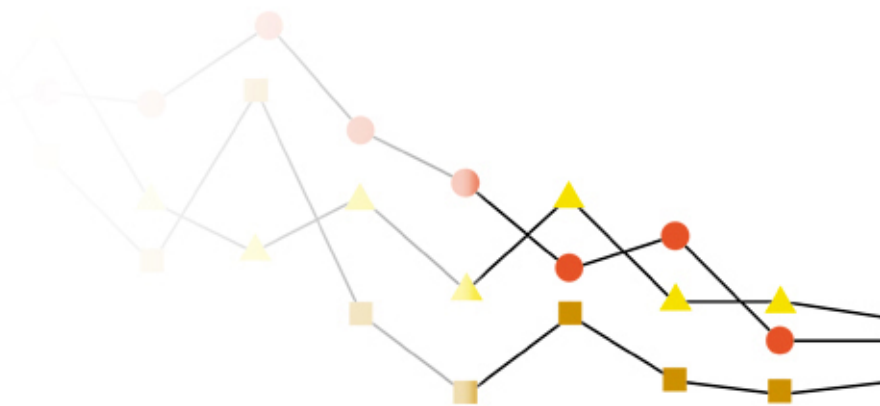
Evaluating Confounding

- Suppose we are measuring the association between an exposure and a disease (eg asbestos and lung cancer)
- We want to control for all potential confounders (eg, age, gender, smoking)
- Ideally we would run
  - A univariate model (asbestos only)
  - A ‘full’ model (all potential confounders and asbestos



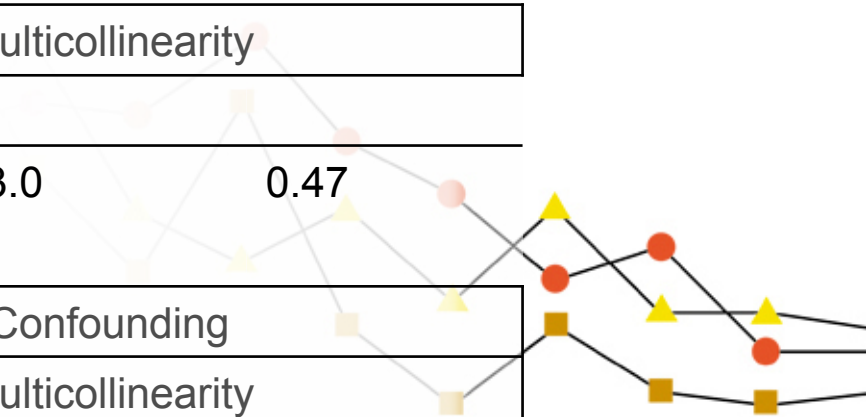
If the RR estimate for asbestos changes when we add the other variables to the model then there was confounding by some or all of these other variables (age, gender, smoking).

Ideally we want to control for all potential confounders and we want to run the “full” model.



Example

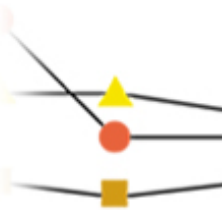
Model	Variables	Asbestos b_1	Asbestos RR	Asbestos $SE(b_1)$
1	Asbestos	0.693	2.0	0.24
2	Asbestos +age +sex +smoking	0.693	2.0	0.24
		No confounding		
2	Asbestos +age +sex +smoking	1.099	3.0	0.25
		Confounding		
2	Asbestos +age +sex +smoking	0.693	2.0	0.47
		Multicollinearity		
2	Asbestos +age +sex +smoking	1.099	3.0	0.47
		Confounding		
		Multicollinearity		



Chapter 9 (additional material)

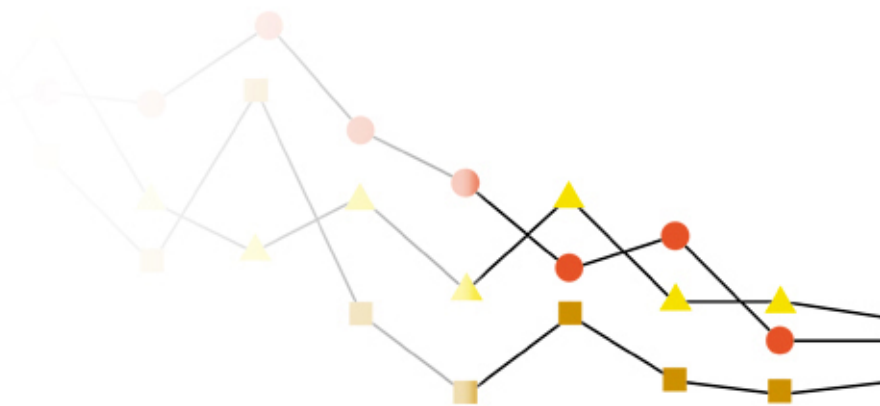
Multiple regression

- Why use multiple regression?
- The basic regression model
- Interaction
- Model selection
- Regression diagnostics
- Approaches to regression



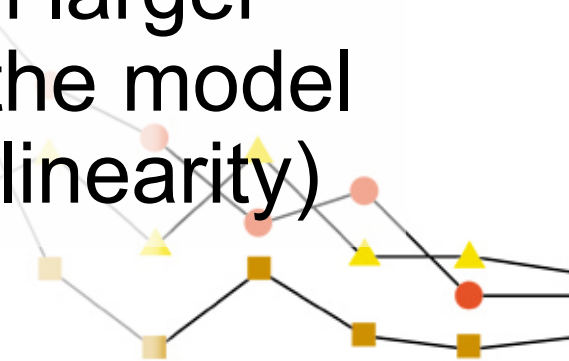
Regression Diagnostics

- Multicollinearity
- Influential data points
- Goodness of fit

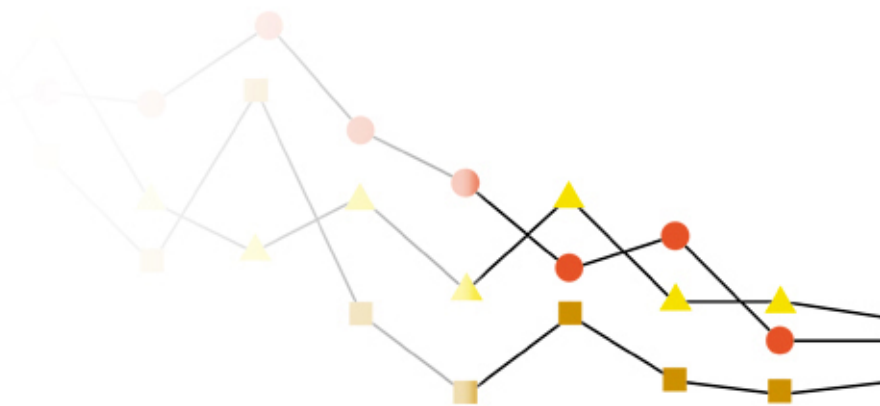


Multicollinearity

The major concern of regression diagnostics is (or should be) the potential problem of *multicollinearity*. This occurs when there is a strong correlation between one or more "confounders" and the main exposure. This will cause the main exposure estimate to be unstable and its SE will become much larger when the "confounder" is included in the model (this is the best way to detect multicollinearity)

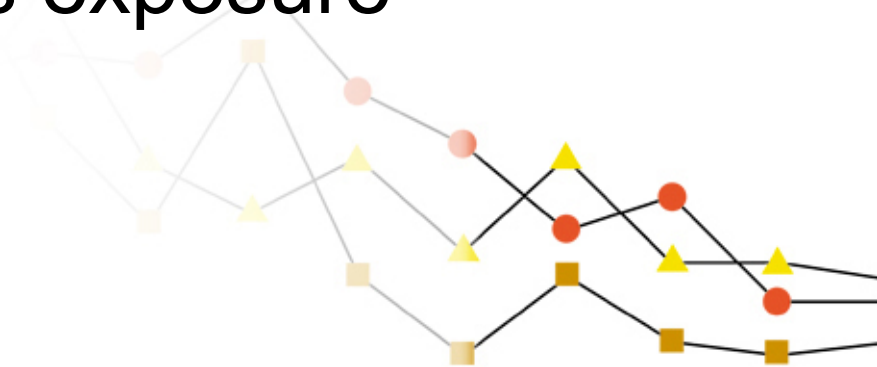


- If the source of multicollinearity is *not* a strong risk factor (and therefore not a strong confounder) then it should *not* be included in the model
- If the source of multicollinearity is a strong risk factor then it should be included in the model and the problem of multicollinearity is insoluble



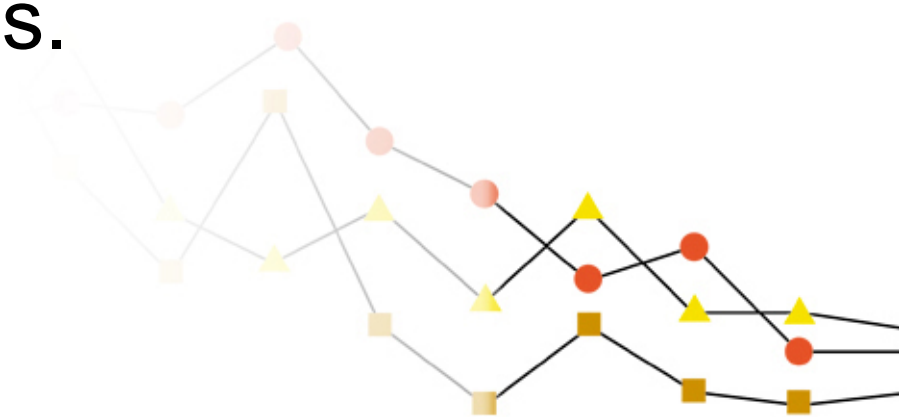
Influential Data Points

- These are data points which strongly influence the maximum likelihood estimates
- For example, if one person with a very heavy exposure lives to be 100, then this will have a big effect on the effect estimate in an analysis using a continuous exposure variable



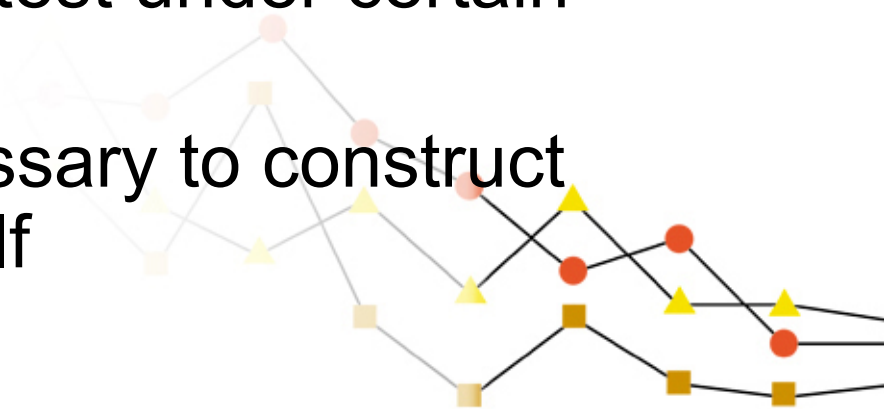
Such points can be identified by deleting each data point in turn to see whether the effect estimate changes substantially.

However, the problem is completely avoided when using categorical rather than continuous exposure variables. This is another reason for using categorical variables.



Goodness of Fit

- Goodness of fit tests involve grouping the data and comparing the observed number of cases in each group with the number predicted by the model
- In Poisson regression the data is already grouped and the model supplies the deviance (which will provide a valid goodness of fit test under certain conditions)
- In logistic regression it is necessary to construct the groups and the test yourself



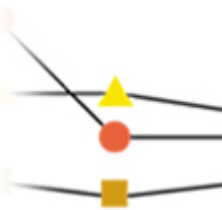
Note

- Goodness of fit tests assess whether the model “predicts the observed data well”. They do not assess confounding of the main exposure variable. It is possible for a model to fit poorly but still estimate the exposure effect correctly
- It is also possible for a model to fit well but still estimate the main exposure effect poorly

Chapter 9 (additional material)

Multiple regression

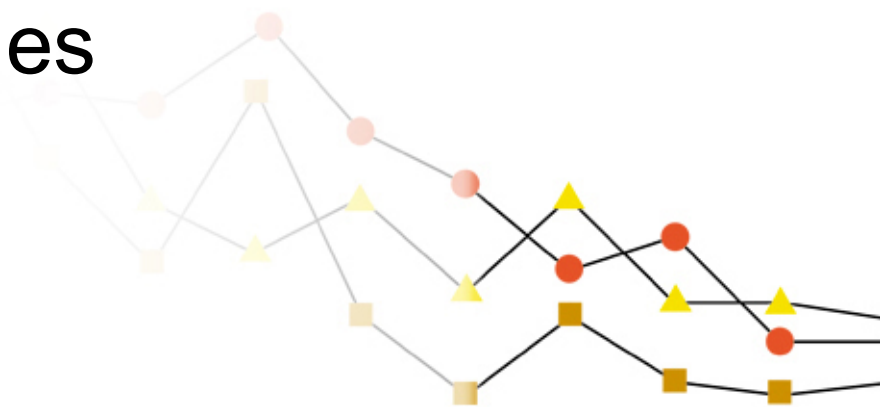
- Why use multiple regression?
- The basic regression model
- Interaction
- Model selection
- Regression diagnostics
- Approaches to regression



Approaches to Regression

"Traditional" *statistical approaches* involve using models for prediction:

- The aim is to achieve a model that *"fits well"*
- The aim is also to achieve a model that is *"parsimonious"* in that it fits well with the minimum number of variables



Approaches to Regression

Thus in “traditional” *statistical approaches* decisions on adding or deleting variables are based on:

- *Statistical significance*
- *Goodness of fit*

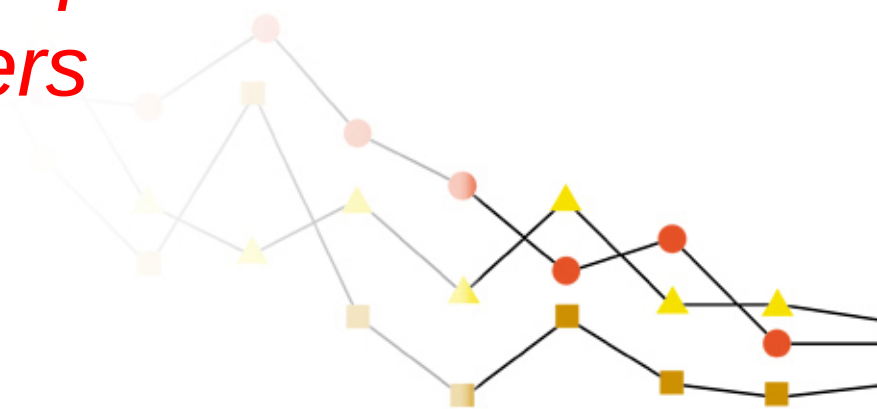
Interaction may be of interest if including interaction terms improves the goodness of fit

Approaches to Regression

Epidemiological approaches involve using models for

- *Effect estimation*
- *Etiologic understanding*

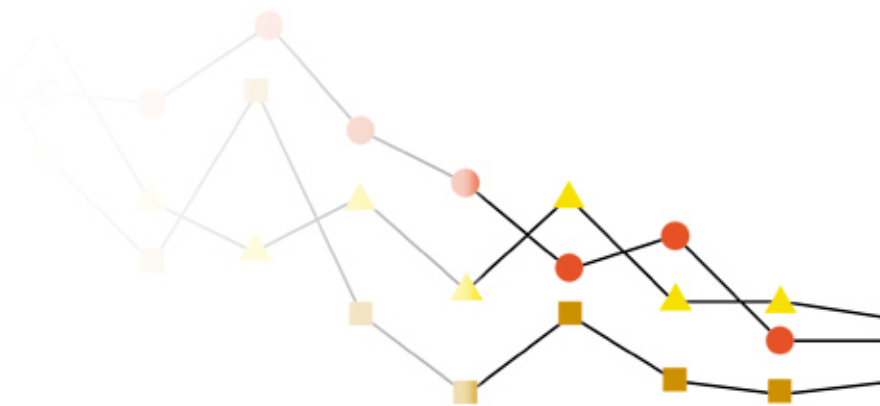
There is usually one main *exposure* and several potential *confounders*



Approaches to Regression

Thus, in *epidemiological approaches*

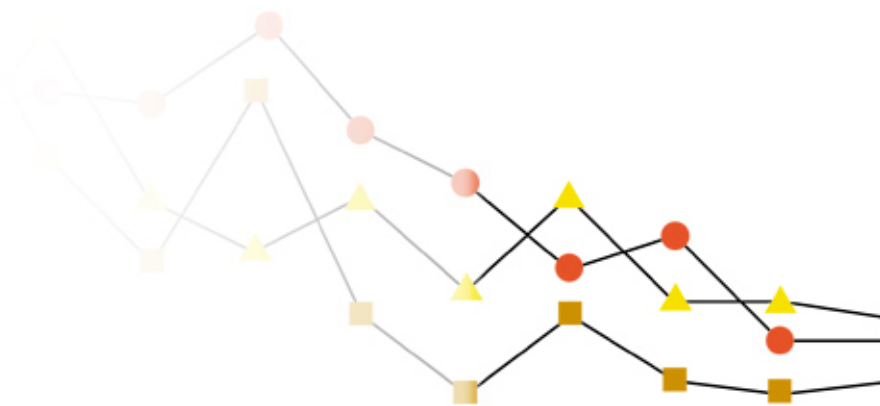
- The main exposure should always be in the model
- Decisions on adding potential confounders should be based on whether the main exposure effect changes



Approaches to Regression

Thus, in *epidemiological approaches*

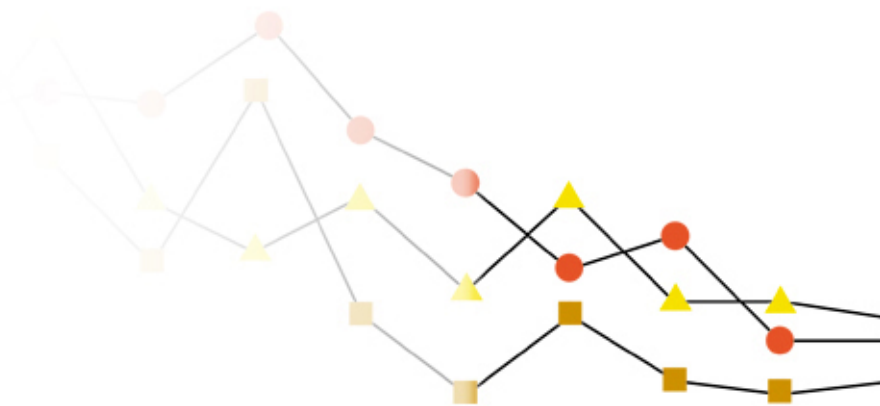
- A variable that "adds significantly" to the model may not be a confounder
- A variable that does not "add significantly" may be a confounder



Approaches to Regression

Thus, in *epidemiological approaches*

- All potential *confounders* should be controlled if possible
- Adding variables that are strongly correlated with exposure will result in *multicollinearity* making the model unstable



Approaches to Regression

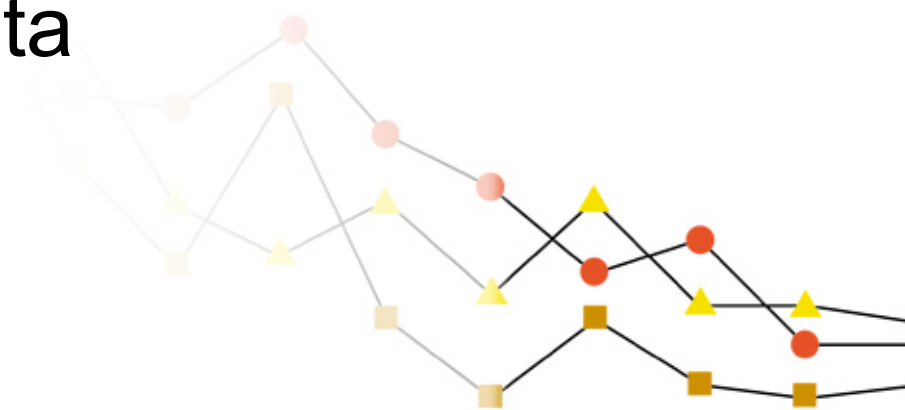
Thus in *epidemiological approaches*: decisions on adding or deleting variables are based on the need to

- Control *confounding*
- Avoid *multicollinearity*

Interaction is of lesser concern unless there are strong a priori to examine it

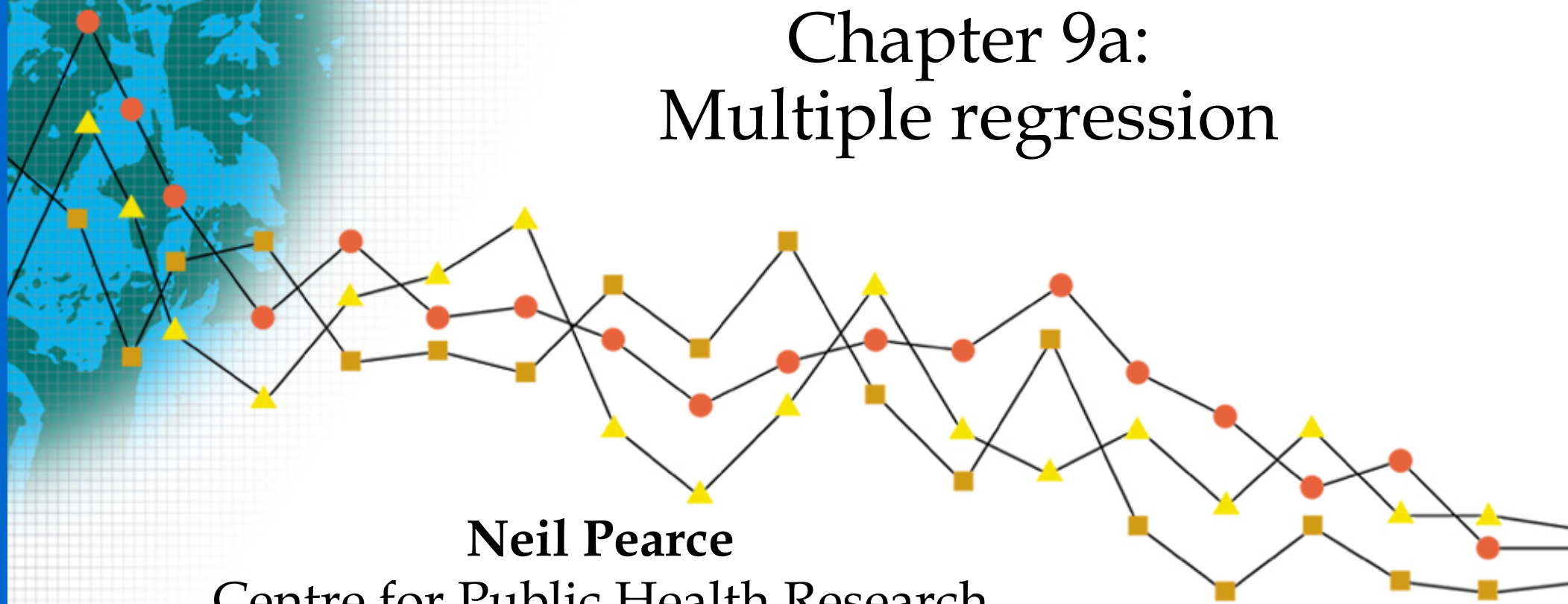
Approaches to Regression

- The most important issue is often to consider the time pattern of exposure and effect
- We may use various deductive etiologic models to summarize exposure information and to assess how well the different exposure models fit the data



A short introduction to epidemiology

Chapter 9a: Multiple regression



Neil Pearce
Centre for Public Health Research
Massey University
Wellington, New Zealand