Structural Equation Models

BIOST 578A, Winter 2007

Ken Rice

January 17, 2007

Overview

- Multivariate data
- **P**rincipal **C**omponents **A**nalysis
- **S**tructural **E**quation **M**odeling - an overview
- Some problems with SEM in practice
- Further references

Slides, further reading, annotated R code, links to other software will be available at `http://courses.washington.edu/bios578b/SEM.html`

Correlations: a quiz!

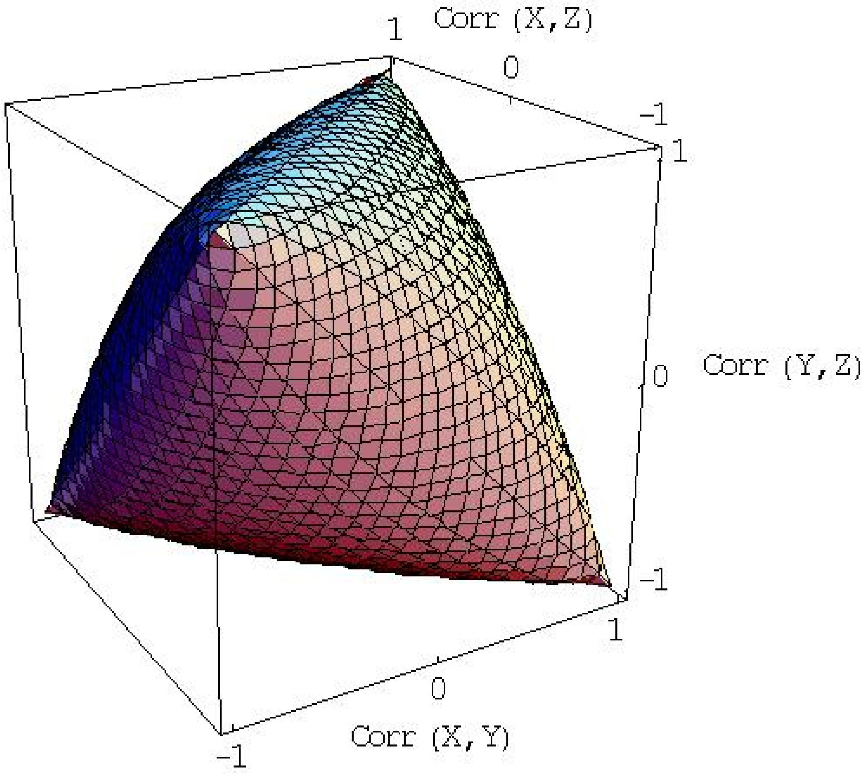
- We will discuss **correlation** between random variables
- For two variables, we **know** $Corr(X, Y)$ is between -1 and 1.
- For *multiple* random variables, the rules are **more complex**

Some examples; what might they represent? Which situation is impossible?

$Corr(X, Y)$	$Corr(Y, Z)$	$Corr(X, Z)$
0	0	0 ▮
0	0	-1 ▮
1	1	1 ▮
-1	-1	-1

Correlations: be careful

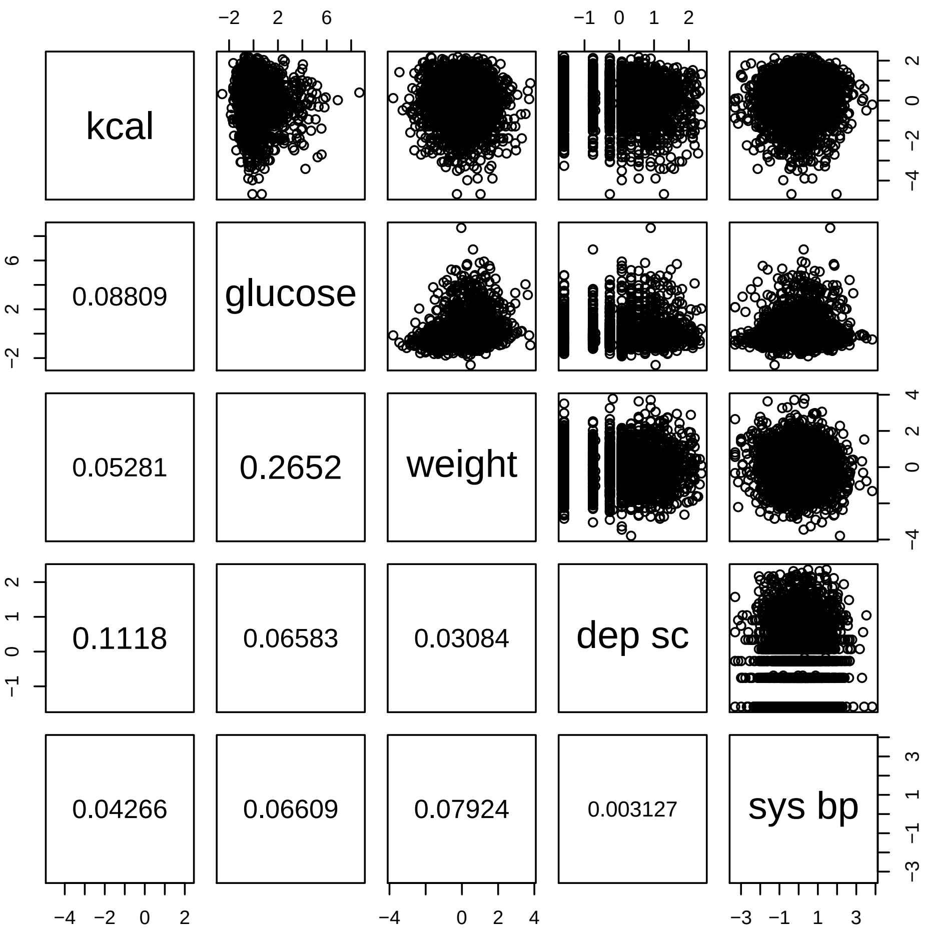
Intuition doesn't go very far in this area



Because of these restrictions, **special techniques** have to be used

Some real data

Standard “epi” data – log transformed, still not quite Normal



Techniques you may know

If we're interested in 'explaining' blood pressure, does `kcal` or `glucose` do a better job?

- For `kcal`, $R^2 = 0.044^2$

- For `glucose`, $R^2 = 0.066^2$

Combining the two covariates (with multiple linear regression), we look for the **linear combination**

$$\beta_1 \texttt{kcal} + \beta_2 \texttt{glucose}$$

which explains most variation in blood pressure (get $R^2 = 0.082^2$)

Techniques you may know

- With no **single** outcome of interest, may still want to explore the **covariation** of the variables

- Similar ideas apply regarding ‘variance explained’; which straight line gets closest to **all** the data?

Variance explained: multivariate

Switch to thinking about the correlation of the whole dataset;

	kcal	glucose	weight	dep sc	sysbp
kcal	**1.000**	-0.088	-0.053	-0.112	0.043
glucose	-0.088	**1.000**	0.265	0.066	0.066
weight	-0.053	0.265	**1.000**	-0.031	-0.079
dep sc	-0.112	0.066	-0.031	**1.000**	-0.003
sys bp	0.043	0.066	-0.079	-0.003	**1.000**

- Different **linear combinations**;

  $$\beta_1\texttt{kcal} + \beta_2\texttt{glucose} + \beta_3\texttt{weight} \ldots$$

  explain different amounts of variation
- Which linear combination explains the **most** variation?
- Tells us which combination of variates **best summarize** what's going on
- More than one correlation = difficult problem (teabag)

Principal Components Analysis (PCA)

- Finds the best combination, 2nd best; the **ordering** is by **how much variation they explain**.

- The measure of ‘how much’ is immune to ‘teabag’ problems

	kcal	glu	weight	dep sc	sys bp	Std dev
Component 1	0.374	-0.657	-0.616	-0.212	0.07	1.15
Component 2	-0.541	-0.171	-0.399	0.718	-0.059	1.04
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Get five components, each contains five **loadings** - the β parameters for each covariate

- **Size** of loading reflects weight of each variable

- Typically see $\beta_1 + \beta_2$, $\beta_1 - \beta_2$ in first two components

Implementing PCA in free software

Example code for the `R` software; (the epi data is called **`sem3`**)

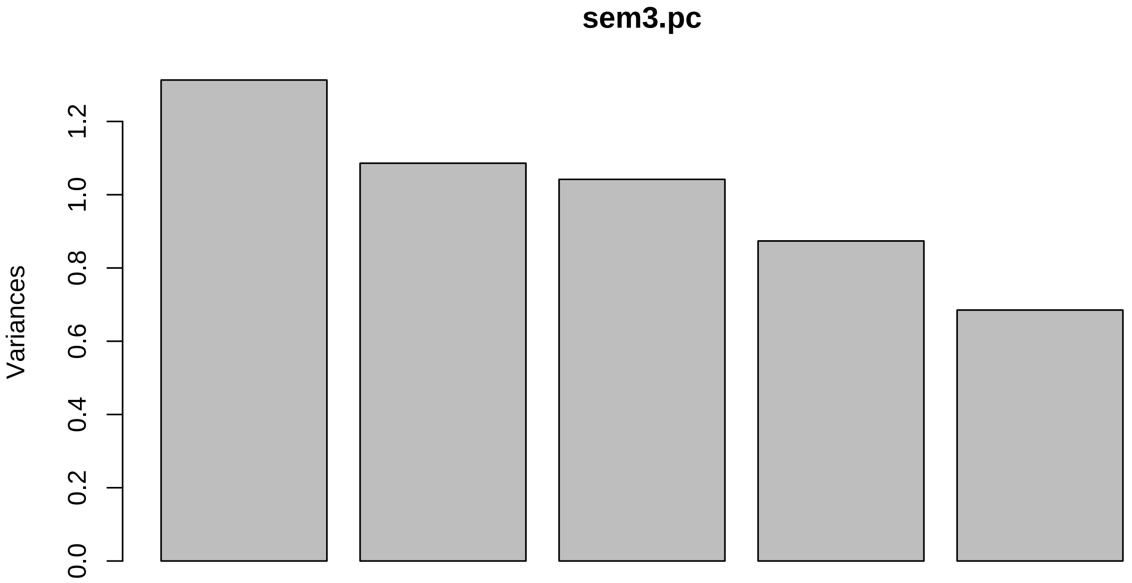
```
> prcomp(sem3)
Standard deviations:
[1] 1.1459222 1.0421414 1.0207497 0.9347535 0.8277135

Rotation:
                 PC1         PC2        PC3         PC4         PC5
kcalbl    0.37372471 -0.54143356  0.1373456 -0.73855553 -0.05339919
glu44    -0.65655992 -0.17137472  0.2916036 -0.10421134 -0.66608342
weightbl -0.61608895 -0.39888917 -0.1746027 -0.09871718  0.64891498
depscrbl -0.21152681  0.71757379  0.1514179 -0.61844572  0.18692717
avsysy11  0.07035432 -0.05878513  0.9179815  0.22683812  0.31216820
```

- ‘Rotations’ also known as principal components, the std deviations tell you about how much variance is explained
- Numerically, not too hard for e.g. 20 covariates
- Also available in other packages (I’m not an expert)
- You **cannot** ‘eyeball’ this sort of thing

How many components is enough?

```
> plot(prcomp(sem3))
```



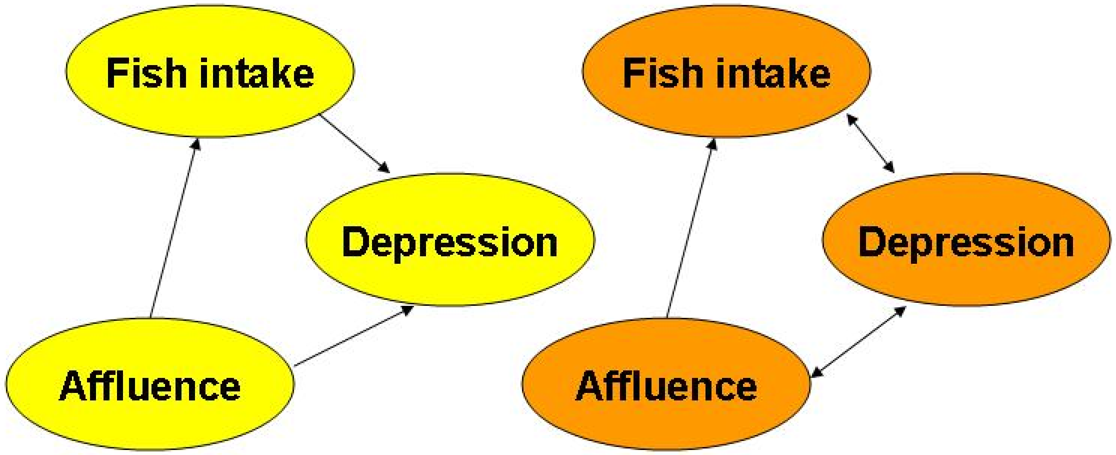
Looking for an ‘elbow’ here... not always present

Some issues with PCA

- Nice, simple, investigative technique; no hypotheses, no p-values
- **Perfectly reasonable** for exploratory work
- Five variables == five components, no more – usually just the first one/two of interest
- **Sign** is arbitrary – multiply each row by -1 to no effect
- All data here was normalized before we began. (Mean=0, Std Dev=1) – in practice variables should all be in the same *sort* of range, or 'large' variables will dominate

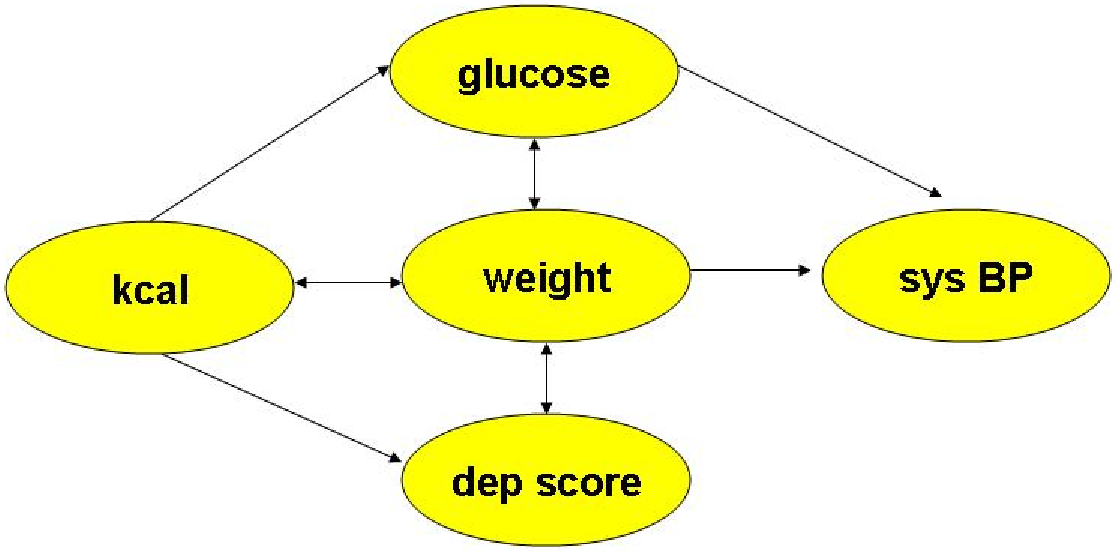
Structural Equation Modelling

- As with PCA, all variables are treated as 'outcomes of interest'
- Impose structure on the **way** they co-vary (hypothesis);



- You can think of each arrow as a linear regression - all effects are causal
- There are 64 potential ways to 'wire up' these variables

Standard epi data again

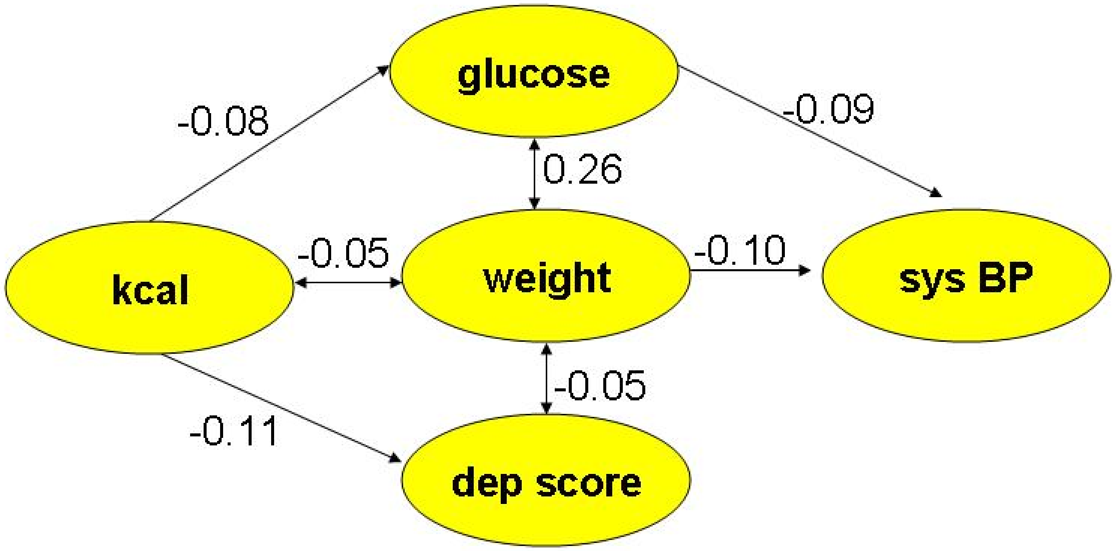


- We will fit β parameters for each line

- Comparing these we get **some** idea of which connections matter most

- All fitting happens at once, better than several linear regression (teabag!)

Output from epi example



- Multiply along pathways to assess their influence (path analysis)

- Likelihood ratio tests let us **compare** different path diagrams (χ^2 tests).

- Each β parameter also has an associated p-value, confidence interval

R code to do SEM

```
library(sem)
sem.model <- matrix(c(
     ’kcalbl    -> glu44’,    ’beta1’, NA,
     ’kcalbl   <-> weightbl’, ’beta2’, NA,
     ’kcalbl    -> depscrbl’, ’beta3’, NA,
     ’kcalbl   <-> kcalbl’,        NA,  1,
     ’glu44    <-> weightbl’, ’beta4’, NA,
     ’glu44     -> avsysy11’, ’beta5’, NA,
     ’weightbl <-> depscrbl’, ’beta6’, NA,
     ’weightbl  -> avsysy11’, ’beta7’, NA,
     ’glu44    <-> glu44’,     ’sig1’, NA,
     ’weightbl <-> weightbl’,  ’sig2’, NA,
     ’depscrbl <-> depscrbl’,  ’sig3’, NA,
     ’avsysy11 <-> avsysy11’,  ’sig4’, NA),
ncol=3, byrow=TRUE)
sem(sem.model, S=cov(sem3), N=length(sem3))
```

Really just writing down everything from the path diagram, and giving it a name

Other software to do this

- `R` is free but only lets you model the covariance - typically we also want means (intercepts) estimated simultaneously

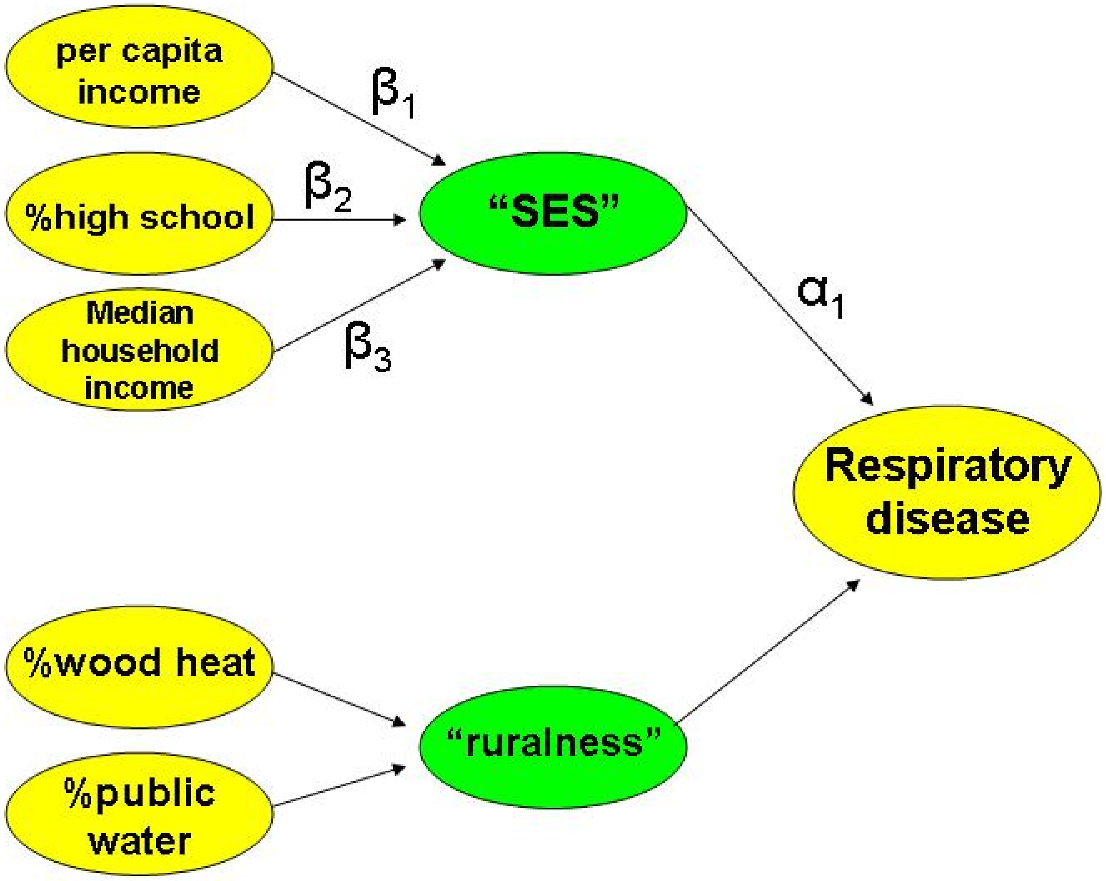
- SAS has PROC CALIS

- Specialized software usually required – LISREL, COSAN, RAM

- Can do much more than R, SAS, but takes a bit of learning

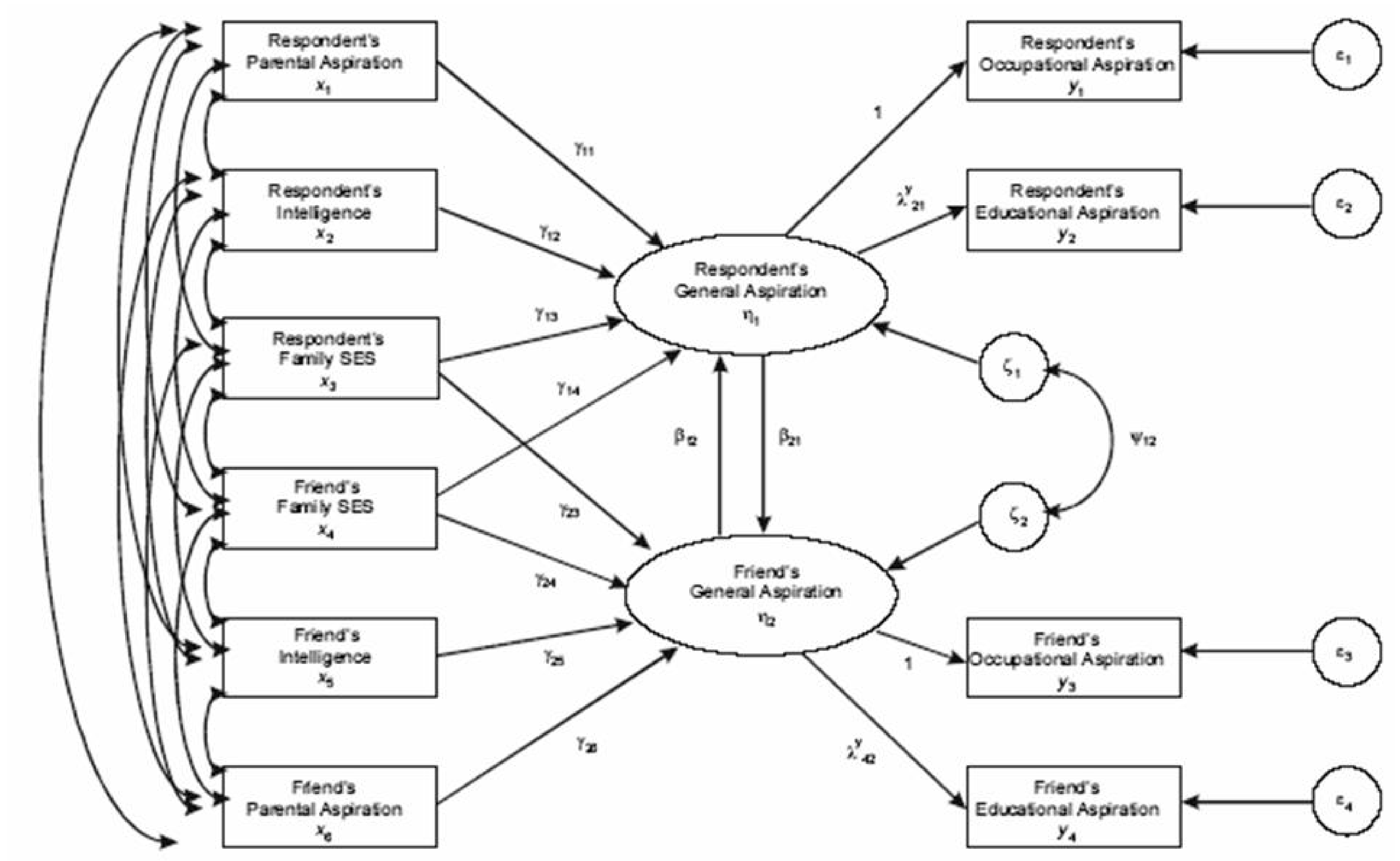
- Natural for Bayesians (!) – who are used to thinking about assumptions

Latent variables



More advanced, "reaching around in the dark" (only find what you assume's there)

A real, “simple” example (!)


Respondent's Parental Aspiration x_1
Respondent's Intelligence x_2
Respondent's Family SES x_3
Friend's Family SES x_4
Friend's Intelligence x_5
Friend's Parental Aspiration x_6
γ_{11}
γ_{12}
γ_{13}
γ_{14}
γ_{23}
γ_{24}
γ_{25}
γ_{26}
Respondent's General Aspiration η_1
Friend's General Aspiration η_2
β_{12}
β_{21}
ζ_1
ζ_2
ψ_{12}
1
λ^y_{21}
1
λ^y_{42}
Respondent's Occupational Aspiration y_1
Respondent's Educational Aspiration y_2
Friend's Occupational Aspiration y_3
Friend's Educational Aspiration y_4
ε_1
ε_2
ε_3
ε_4

Genetics example: Mendel

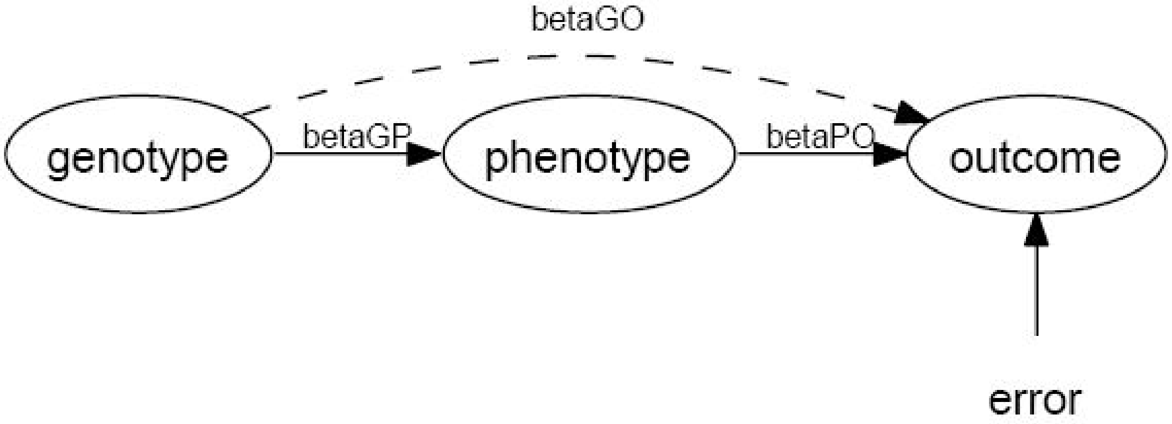
- Geneticists are **very good** at finding genes which control the production of e.g. HDL, LDL, CRP, TLAs

- TLAs (phenotypes) can be associated with **outcomes**; but only because their gene was associated in some other pathway ("route")

- We know genes are not changed by phenotype/outcome, so some arrows are easy

- SEM looks like it could be helpful here

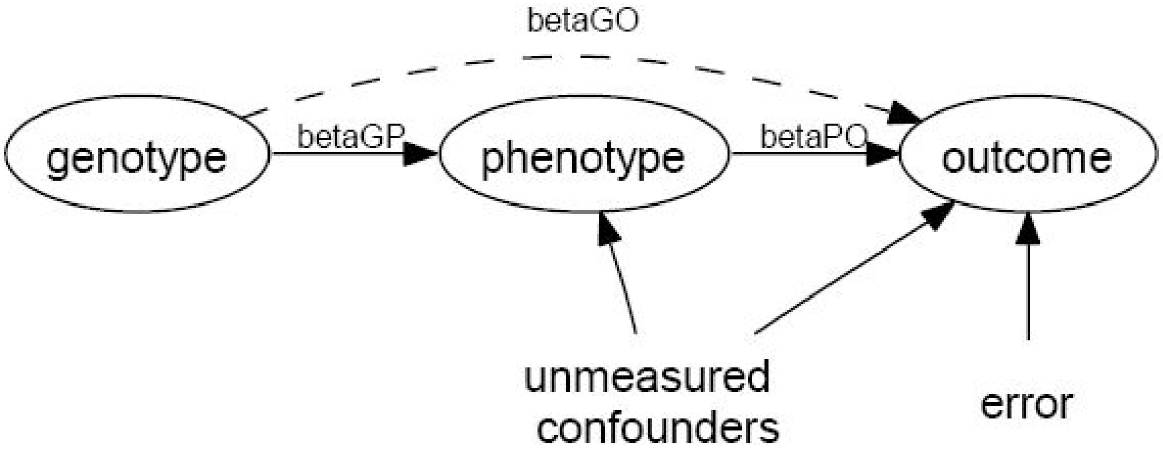
Genetics example: Mendel



SEM ‘adjusts’ $\widehat{\beta}_{PO}$; you get $\widehat{\beta}_{GO}/\widehat{\beta}GP$, estimating a **causal** effect in a valid way (“**M**endelian **R**andomization”)

A **great idea!** (...if life was this simple)

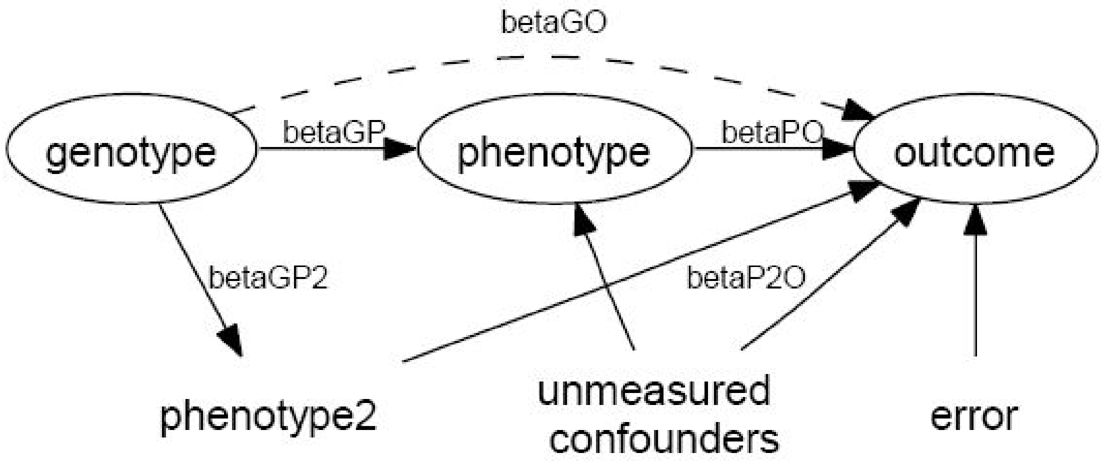
Genetics example: Mendel



Other effects may act on phenotype and outcome. If these are stronger, the MR power is hit hard; e.g. sample size 1000 gives 20% power.

An example of a “weak instrument”

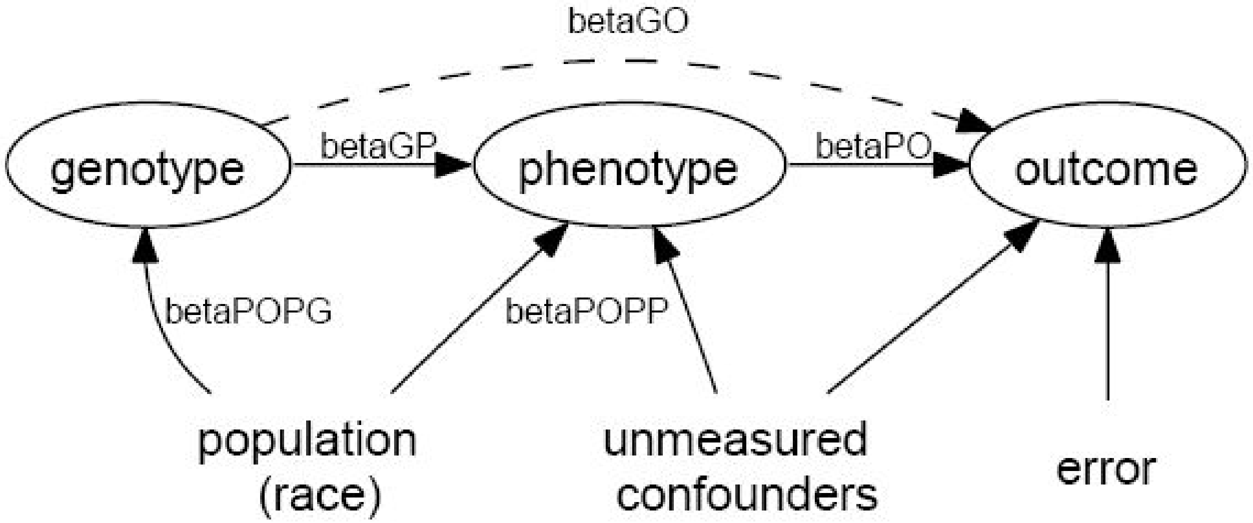
Genetics example: Mendel



If the gene acts in several pathways (“pleiotropy”) our β_{GO} estimate is wrong; $p = 0.05$ has **nothing like** the usual meaning.

Very hard to rule out this situation; you must understand the **whole** system, and have enough data to estimate it well.

Genetics example: Mendel



One version of "population stratification" is shown here; in this format it can actually improve a weak instrument. But have we missed any arrows?

Some problems # 1

- Normality of **all** data is a strong and important assumption
- The shape of the teabag is important but non-intuitive
- You cannot fit all the models you would like, or even the obvious ones (*cf* complete confounding)
- Adjusting the arrows **after** getting the data invalidates **everything**
- Answers depend **strongly** on the way you ‘wired up’ the initial diagram
- There are typically *millions* of ways to do this - note that missing arrows are important
- Estimating some $\beta = 0$ will depend a lot on **how well it was measured**

Some problems # 2

- Not designed for typical applications! (Comes from economics)
- Meant for answering *extremely* precise questions
- You need a *lot* of data, even if the conditions are met
- Practioners have a healthy skepticism for output
- Latent constructions depend *entirely* on the modelling assumptions
- I have avoided a vast number of Greek letters and confusing terminology
- Statisticians feel they would be more productive doing something else

Summary

- Use PCA if it seems helpful (it can be)

- Interpret any path analyses in ‘ball park’ terms

- Avoid full-on SEM if possible

- Avoid using cross-sectional data to try to learn about causality (*cf* ‘Mendelian Randomization’)

Further references

- Venables, WN and Ripley, BD (1997) Modern Applied Statistics with S-PLUS, Springer-Verlag.

- Bollen, KA (1989) Structural Equations with Latent Variables. Wiley

- Fox, J (2002) An *R* and *S-Plus* Companion to Applied Regression. Sage

- Wall, MM and Li, R (2003) Comparison of multiple regression to two latent variable techniques for estimation and prediction. Statistics in Medicine 22:3671–3685

- Feldman, PJ and Steptoe, A (2004) How neighbourhoods and physical functioning are related. Ann Behav Med 27(2):91–99

- Brannick, M. Lecture notes on SEM, link on website