Introduction to Graphical Models with Applications to Quantitative Genetics and Genomics

Armidale - Australia

Jan 29 – Feb 1, 2019

University of New England

Instructors:

- Guilherme J. M. Rosa (Gee-Ler-Me)

  Department of Animal Sciences

  Department of Biostatistics & Medical Informatics

  The University of Wisconsin Madison

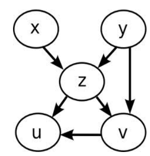
- Francisco Peñagaricano ("Pancho")

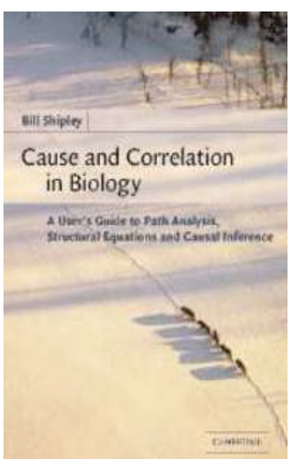
  Department of Animal Sciences

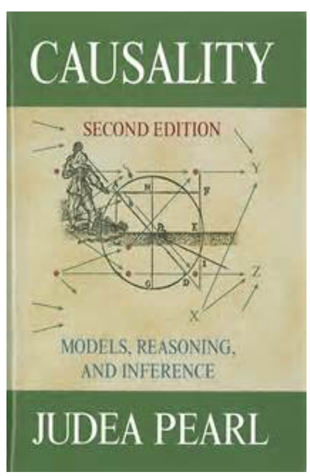
  UF University of Florida

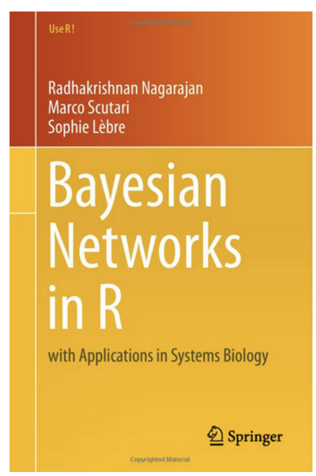
OUTLINE

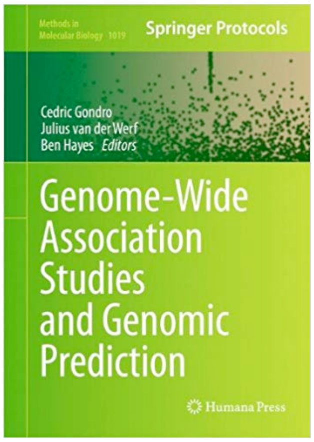
- Correlation and Causation
- Basics of Matrix Algebra, Probability, Random Variables
- Path Analysis
- Test for Independence
- Correlation Networks
- Structural Equation Models in Quantitative Genetics
- Latent Variables
- Bayesian Networks
- GWAS and QTL Analysis

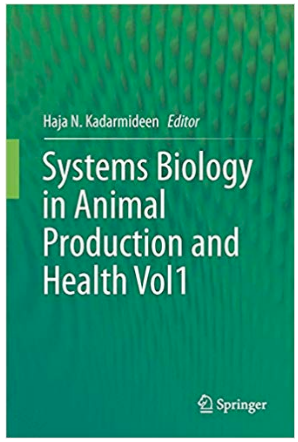


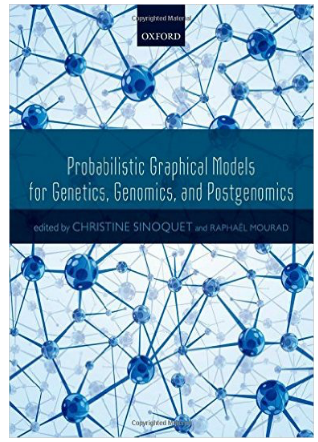












Software

bnlearn
deal
pcalg
catnet

The TETRAD Project

Causal Models and Statistical Data

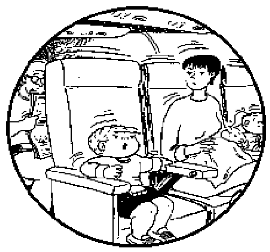
GMTK

The Graphical Models Toolkit

DGM lib

Direct Graphical Models

Correlation & Causation



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

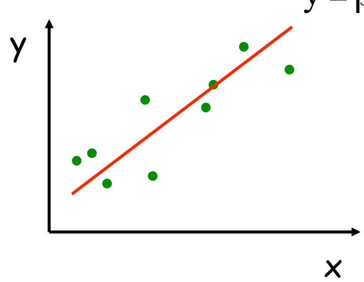
Simple Linear Regression

Data pairs

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

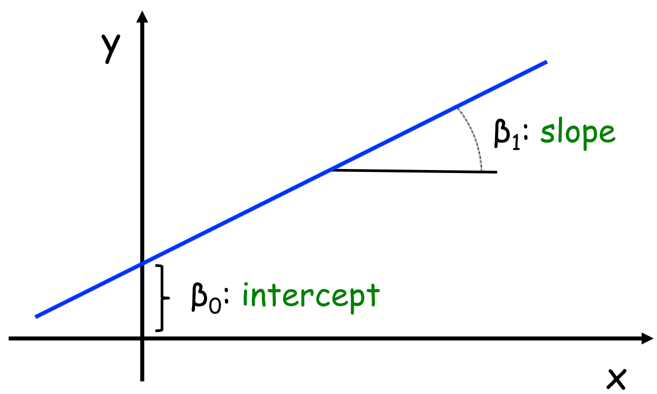
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



➔ β_0 is the intercept; β_1 is the slope

Simple Linear Regression

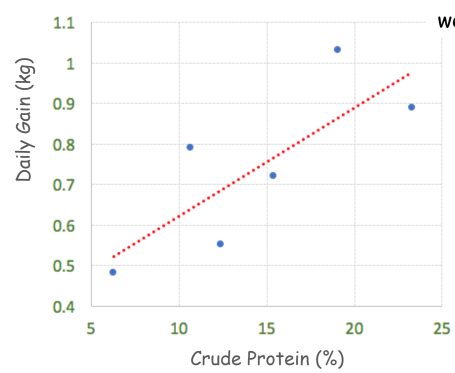
$$E[y] = \beta_0 + \beta_1 x$$



Example: Forage crude protein (% of dry matter) and beef cattle average daily weight gain (kg)

weight gain

CP (%)	DG (kg)
6.3	0.48
10.7	0.79
12.4	0.55
15.4	0.72
19.1	1.03
23.3	0.89



- Estimated regression: DG = 0.3534 + 0.0268 x CP
- What is the interpretation of the regression coefficient (slope)?

Association vs. Causation

z ------ y

"z is associated with y"

z → y

"z causes y"

alcohol and drunkenness

z ← y

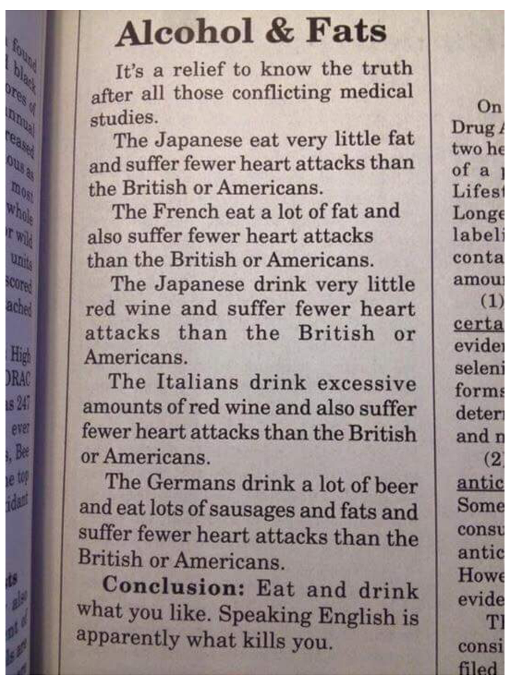
"y causes z"

police officers and criminality

x → z, x → y

"x causes z & y"

water consumption and dehydration



Alcohol & Fats

It's a relief to know the truth after all those conflicting medical studies.

The Japanese eat very little fat and suffer fewer heart attacks than the British or Americans.

The French eat a lot of fat and also suffer fewer heart attacks than the British or Americans.

The Japanese drink very little red wine and suffer fewer heart attacks than the British or Americans.

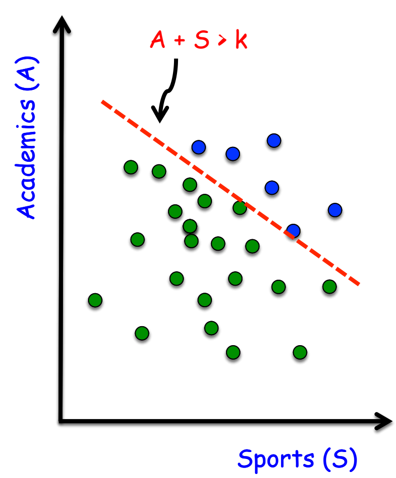
The Italians drink excessive amounts of red wine and also suffer fewer heart attacks than the British or Americans.

The Germans drink a lot of beer and eat lots of sausages and fats and suffer fewer heart attacks than the British or Americans.

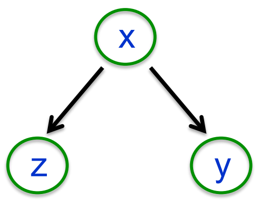
Conclusion: Eat and drink what you like. Speaking English is apparently what kills you.

Confounders

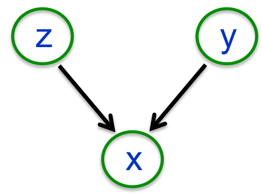
Selection Bias



Confounding and Selection Bias



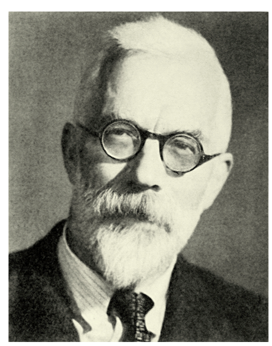
Confounding
(x is a common cause for z and y)



Selection Bias
(z and y observed only
for a subset of x values)

Randomized Trials

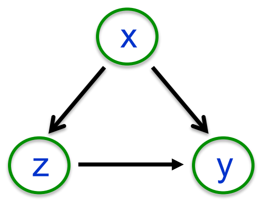
Lady tasting tea



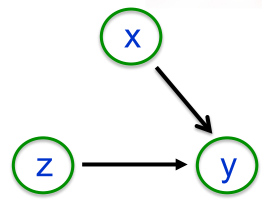
Sir R. A. Fisher

Randomized Experiments

⇨ Testing the effect of z on y.



Causal relationship
between variables



Effect of randomization
applied to variable z

Analysis of Variance (ANOVA)

Group			
G_1	G_2	…	G_k
y_{11}	y_{21}		y_{k1}
y_{12}	y_{22}	…	y_{k2}
⋮	⋮		⋮
y_{1n_1}	y_{2n_2}	…	y_{kn_k}

Sample variance:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$$

$$(N = n_1 + n_2 + \ldots + n_k)$$

y_{ij} : observation on individual (unit) j of group i,
where i = 1,…,k and j = 1,…,n_i

Partitioning Sums of Squares

A fundamental principle of least squares (LS) methods is that variation on a response variable can be partitioned (i.e. divided into parts) according to the sources of the variation. For example, for a 1-way classification model, we have:

Total (corrected) Sum of Squares: $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2 = \underbrace{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}_{\text{Group (between) SS}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}_{\text{Residual (within) SS}}$$

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C \qquad SS_B = \sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n_i} - C \qquad SS_R = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n_i}$$

where: $C = \frac{y_{\bullet\bullet}^2}{N}$ (correction), $y_{i\bullet} = \sum_{j=1}^{n_i} y_{ij}$ and $y_{\bullet\bullet} = \sum_{i=1}^{k} y_{i\bullet}$

Analysis of Variance (ANOVA)

For a statistical comparison of the groups, the following approach can be used to test the null hypothesis (H_0: $\mu_1 = \ldots = \mu_k$) against an alternative hypothesis that there is at least one difference among the group means.

SV	DF	SS	MS	E[MS]	F
Groups	k - 1	SS_B	MS_B	$\sigma^2 + \phi_B$	MS_B/MS_R
Residual	N - k	SS_R	MS_R	σ^2	
Total	N - 1	SS_T	---		

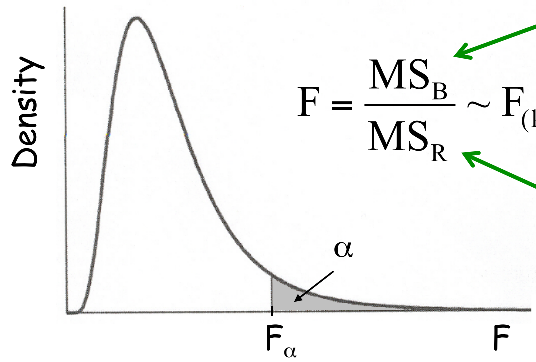
where: SV = Sources of Variation, DF = Degrees of Freedom, SS = Sum of Squares, MS = Mean Squares, E[.] = Expectation, and F is an MS ratio.

Moreover: $\phi_B = \frac{1}{(k-1)}\sum_{i=1}^{k} n_i(\mu_i - \mu)^2$ is a quadratic form involving μ_i's ,

$$MS_B = \frac{SS_B}{(k-1)}, \quad MS_R = \frac{SS_R}{(n-k)}, \quad N = \sum_{i=1}^{k} n_i \text{ and } \mu = \frac{1}{k}\sum_{i=1}^{k}\mu_i$$

Analysis of Variance (ANOVA)

- Assuming normality, i.e. $y_{ij} \sim N(\mu_i, \sigma^2)$, it can be shown that under the null hypothesis the F statistics has an F (Snedecor) distribution as following:



$$F = \frac{MS_B}{MS_R} \sim F_{(k-1, N-k)}$$

between-group variation

within-group variation

Example

Suppose three groups of beef cattle, each fed with a different diet. The results in terms of weight gain are given below :

Diets		
A	B	C
106	84	92
99	99	99
97	89	85
104	80	91
99	82	89
105		92
95		

Model: $\boxed{y_{ij} = \mu_i + e_{ij}}$

- y_{ij}: weight gain observed on animal j of diet i
- μ_i: mean of diet i
- e_{ij}: residual term

- $i = 1, 2, 3$ (Diets A, B and C)
- $j = 1, 2, \ldots, n_i$
- $(n_1 = 7, n_2 = 5, n_3 = 6)$

Example

diets

Sample Means:

Diet		
A	B	C
$y_{1\cdot} = 705$	$y_{2\cdot} = 434$	$y_{3\cdot} = 548$
$y_{\cdot\cdot} = 1687$		

ANOVA Table:

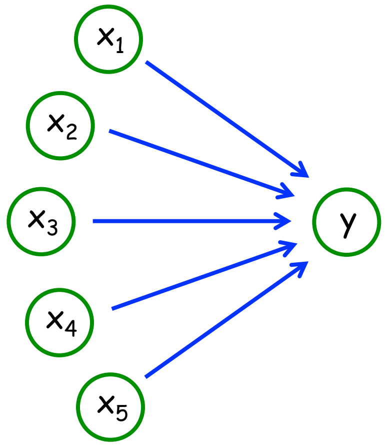
SV	DF	SS	MS	F (p-value)
Diet	2	616.0	308.0	10.37 (0.0015)
Residual	15	445.6	29.7	
Total	17	1061.6		

Observational Studies

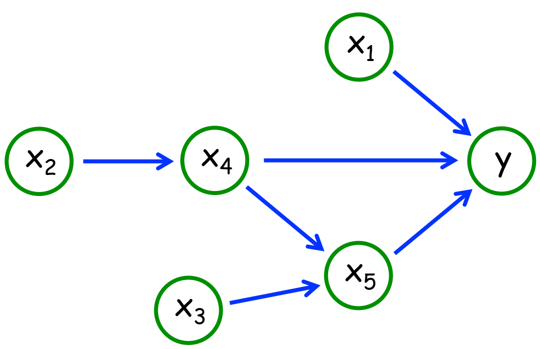
- Lack of randomization due to legal, ethical, or logistics reasons
- Potential bias and confounding effects
- Example: Parenthood and life expectancy

Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i$$



Network Approach



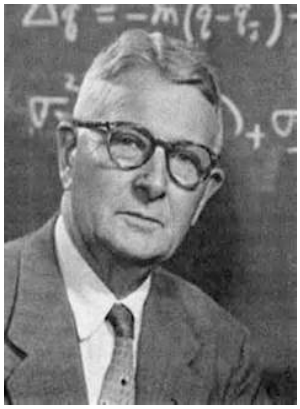
Indirect effects
Direct effects
Total effects

$$y = \beta_0^{(y)} + \beta_1^{(y)} x_1 + \beta_4^{(y)} x_4 + \beta_5^{(y)} x_5 + e^{(y)}$$

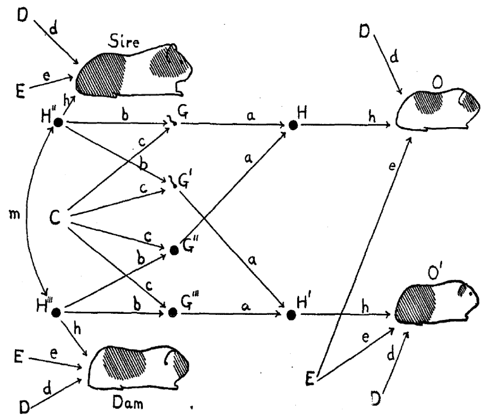
$$x_4 = \beta_0^{(4)} + \beta_2^{(4)} x_2 + e^{(4)}$$

$$x_5 = \beta_0^{(5)} + \beta_3^{(5)} x_3 + \beta_4^{(5)} x_4 + e^{(5)}$$

Path Analysis



Sewall Wright
(1889-1988)



Bayesian Networks