Predictive Analytics

©

Dr Diarmuid McDonnell & Prof. Vernon Gayle

AQMEN

University of Edinburgh

March 2019

Welcome

Why are we here today?

U.K. Industrial Strategy aims to utilise big data to improve economic performance and increase productivity.

Major barrier is the lack of a suitably trained workforce.

U.K. social science stakeholders (e.g. ESRC, Nuffield) believe this discipline can make a major contribution to Industrial Strategy.

Why this type of training?

“Many organizations can barely find a way to use their R/Python programmers on reasonable datasets.”

“The piece missing from the data science movement right now is really simple: intelligent application of data science tools.”

https://www.linkedin.com/pulse/data-science-dead-5-years-less-justin-b-dickerson-phd-mba-pstat- accessed 16.07.2018.

Administrative Data Research Centre Scotland
An ESRC Data Investment
Better Knowledge
Better Society
Supporting researchers to use administrative data for economic and social research
Enabling studies that benefit the public
Helping inform and improve policy-making
W. adrn.ac.uk
E. adrc-s@ed.ac.uk
AQMeN
Applied Quantitative Methods Network

What is the social science contribution?

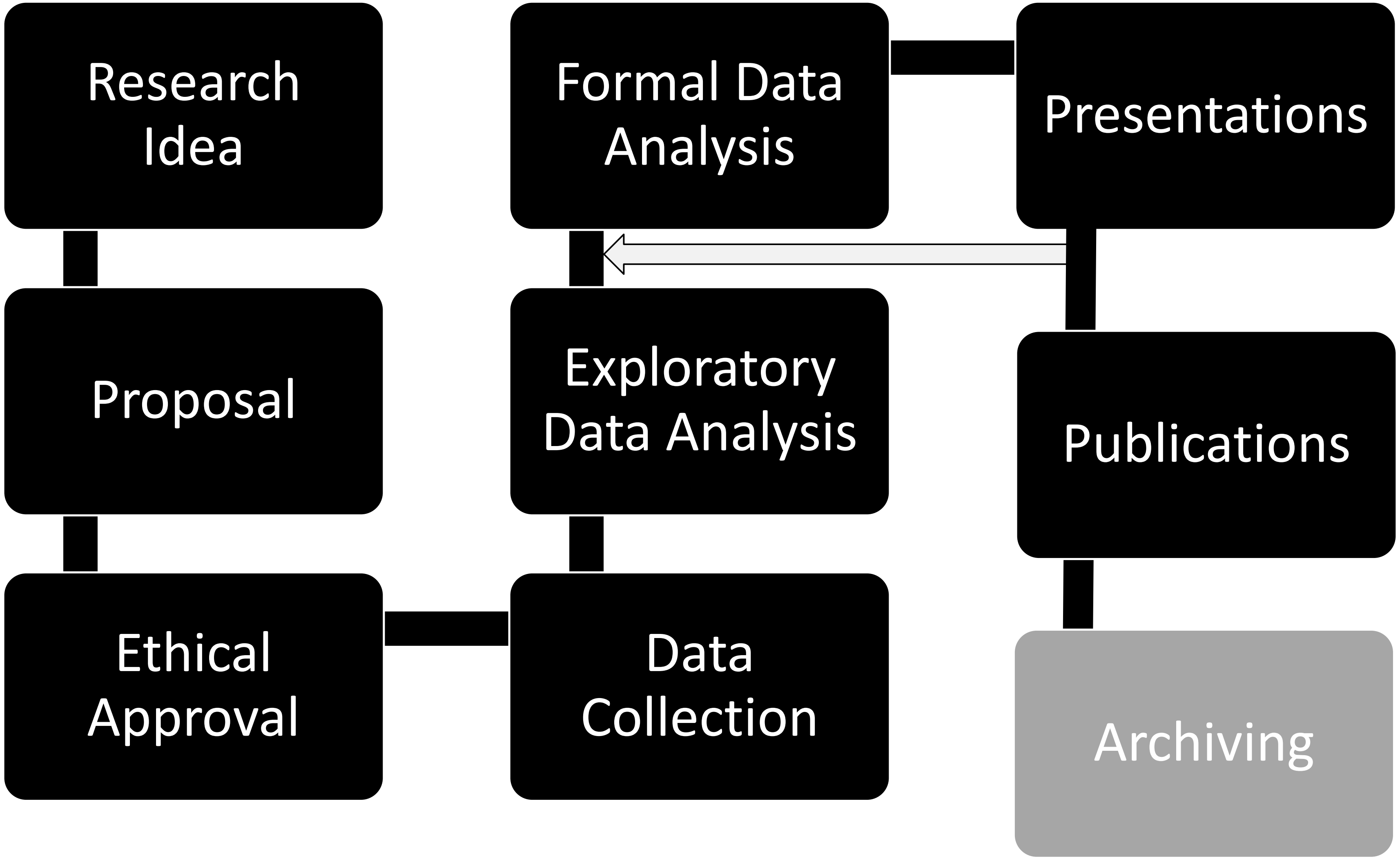
Data Science ≠ Computer Science

Big Data ≈ Small Data

QM Social Scientists = Data Literate

The Workflow

The Workflow



Academic / Commercial Life

1. Measured by outputs

2. Structured by deadlines

Organised workflow helps to make progress

ALL SERIOUS WORK MUST BE REPRODUCIBLE!

There MUST be an audit trail

Documentation

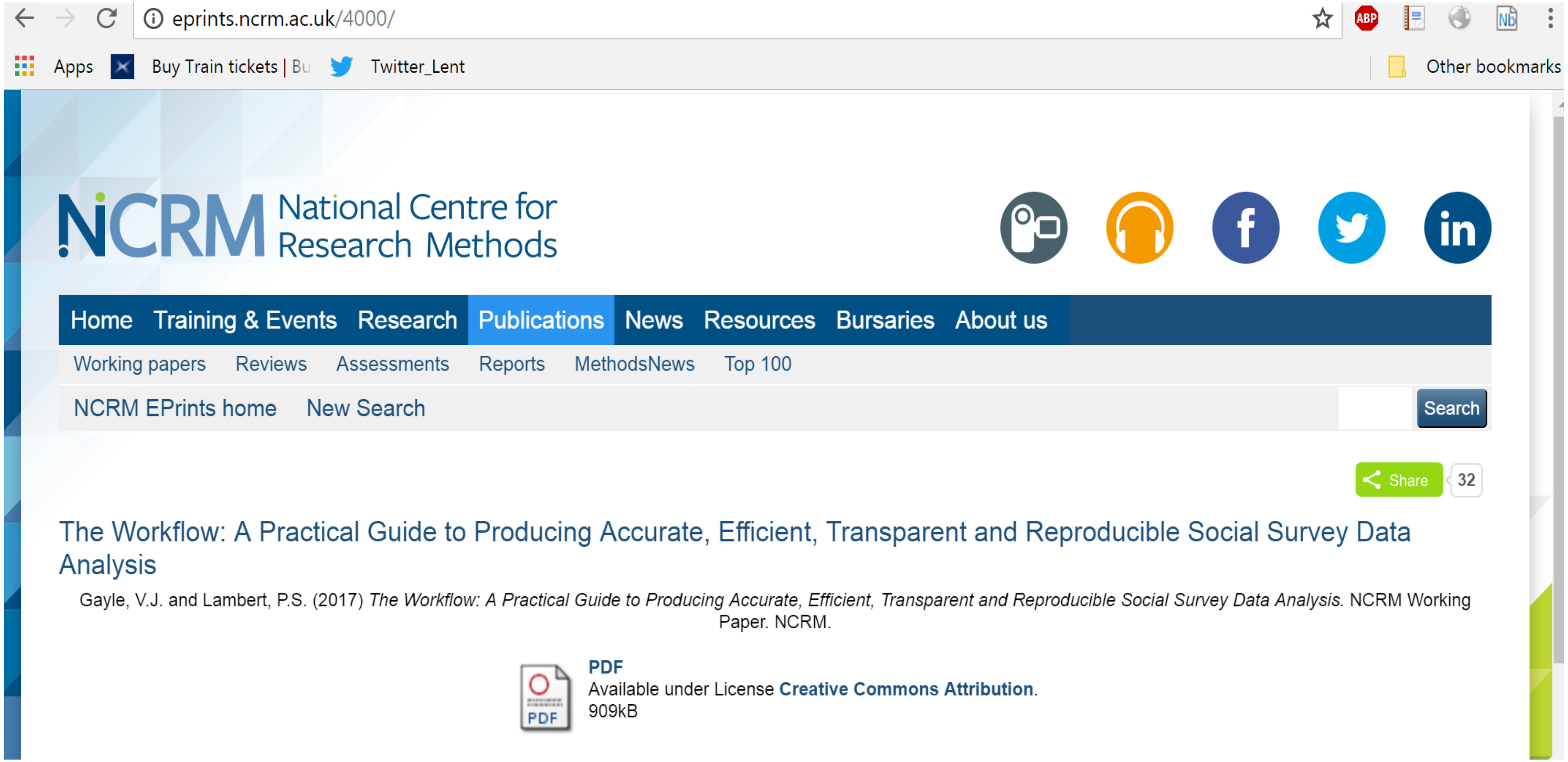
Consistent Working Practices

Standard and routinized ways of doing things

- Style of scripts (.do .sps files etc.)
- Directory structures
- File names
- Variable names
- Variable labels

Estimating Work Time…

http://eprints.ncrm.ac.uk/4000/



The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis

Gayle, V.J. and Lambert, P.S. (2017) *The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis.* NCRM Working Paper. NCRM.

PDF
Available under License **Creative Commons Attribution**.
909kB

J. Scott Long has posted a really good pdf version of a talk on the workflow

http://www.ihrp.uic.edu/files/Workflow%20Slides%20JSLong%20110410.pdf

Thinking Predictively

(Statistical Models of Predictive Analytics for Social Research)

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

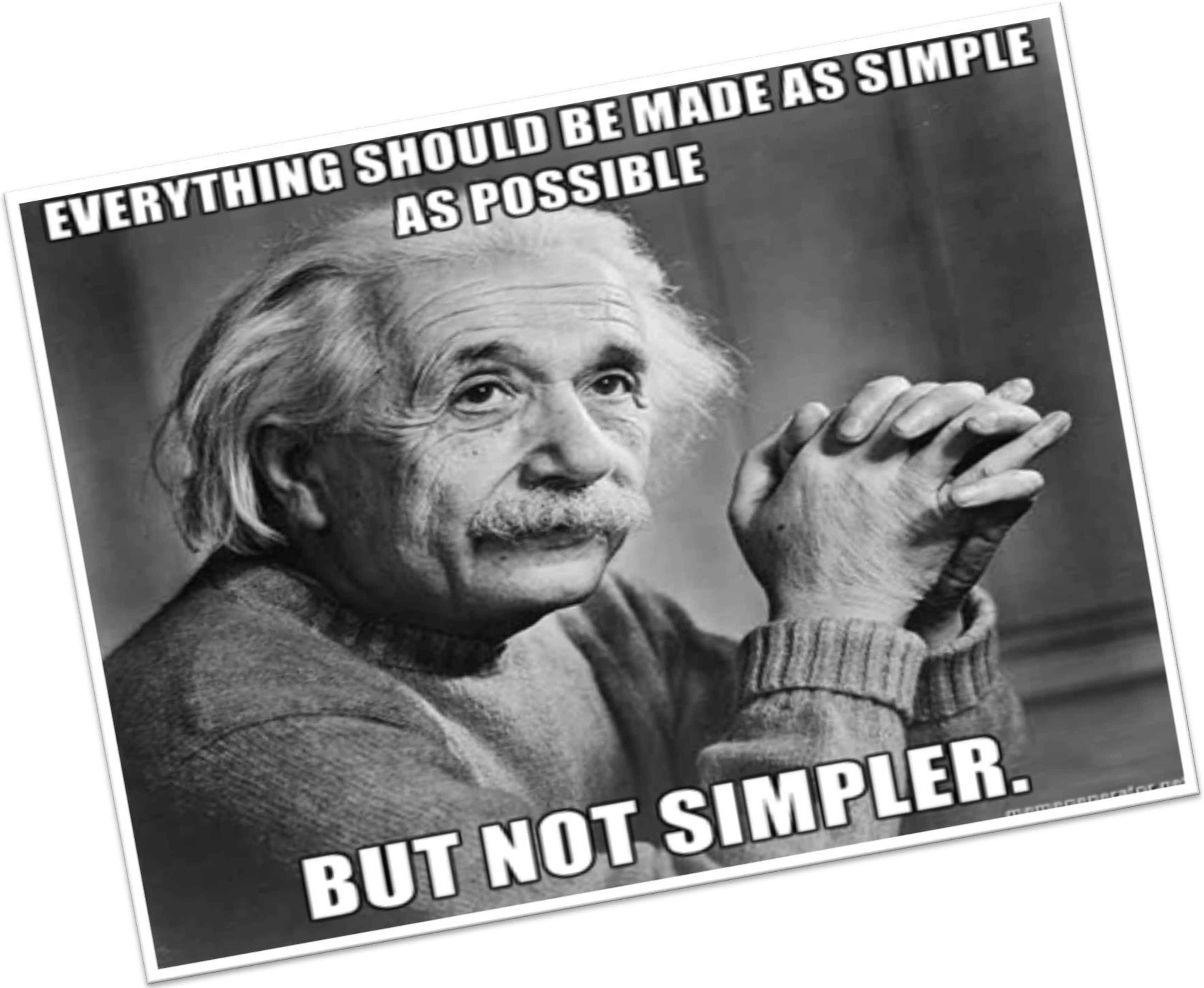
vernon.gayle@ed.ac.uk
@profbigvern

2019

Thinking Predictively

(Statistical Models/Predictive Analytics for Social Research)

First Aid
Made Easy
VERNON


EVERYTHING SHOULD BE MADE AS SIMPLE
AS POSSIBLE
BUT NOT SIMPLER.

International Journal of Sport Nutrition and Exercise Metabolism, 2005, 15, 386-400

Effects of Acute Versus Chronic L-Carnitine L-tartrate Supplementation on Metabolic Responses to Steady State Exercise in Males and Females

Weronika N. Abramowicz and Stuart D.R. Galloway

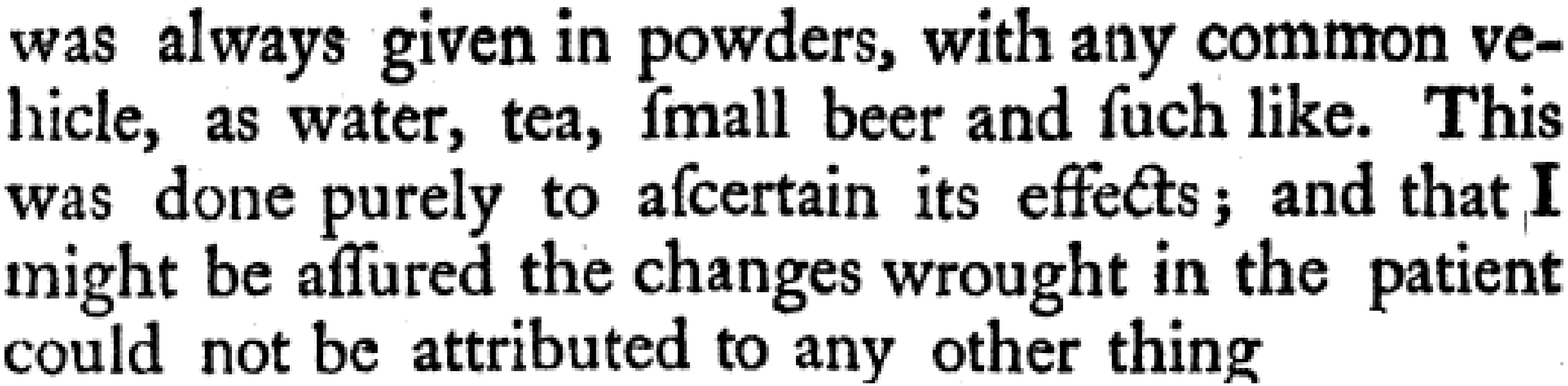
Twelve healthy active subjects (6 male, 6 female) performed 60 min of exercise (60% VO_{2max}) on 3 occasions after supplementing with L-Carnitine L-tartrate (LCLT) or placebo. Each subject received a chronic dose, an acute dose, and placebo in a randomized, double-blind crossover design. Dietary intake and exercise were replicated for 2 d prior to each trial. In males there was a significant difference in rate of carbohydrate (CHO) oxidation between placebo and chronic trials ($P = 0.02$) but not placebo and acute trials ($P = 0.70$), and total CHO oxidation was greater following chronic supplementation vs. placebo (mean ± standard deviation) of 93.8 (17.3) g/hr and 78.2 (23.3) g/h, respectively). In females, no difference in rate of, or total, CHO oxidation was observed between trials. No effects on fat oxidation or hematological responses were noted in either gender group. Under these experimental conditions, chronic LCLT supplementation increased CHO oxidation in males during exercise but this was not observed in females

Key Words: carbohydrate oxidation, fat oxidation, gender

Rev Edward Stone (1702-1768)

Discovered the active ingredient of aspirin

He wrote to the Royal Society on 25 April 1763



was always given in powders, with any common vehicle, as water, tea, ſmall beer and ſuch like. This was done purely to aſcertain its effects; and that I might be aſſured the changes wrought in the patient could not be attributed to any other thing

I have no other motives for publiſhing this valuable ſpecific, than that it may have a fair and full trial in all its variety of circumſtances and ſituations, and that the world may reap the benefits accruing from it. For theſe purpoſes I have given this long and minute account of it, and which I would not have troubled your Lordſhip with, was I not fully perſuaded of the wonderful efficacy of this Cortex Salignus in agues and intermitting caſes, and did I not think, that this perſuaſion was ſufficiently ſupported by the manifold experience, which I have had of it.

I am, my Lord,

with the profoundeſt ſubmiſſion and reſpect,

your Lordſhip's moſt obedient
humble Servant

Chipping-Norton,
Oxfordſhire,
April 25, 1763.

Edward Stone.

BRITISH MEDICAL JOURNAL

LONDON SATURDAY OCTOBER 30 1948

STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS

A MEDICAL RESEARCH COUNCIL INVESTIGATION

The following gives the short-term results of a controlled investigation into the effects of streptomycin on one type of pulmonary tuberculosis. The inquiry was planned and directed by the Streptomycin in Tuberculosis Trials Committee, composed of the following members: Dr. Geoffrey Marshall (chairman), Professor J. W. S. Blacklock, Professor C. Cameron, Professor N. B. Capon, Dr. R. Cruickshank, Professor J. H. Gaddum, Dr. F. R. G. Heaf, Professor A. Bradford Hill, Dr. L. E. Houghton, Dr. J. Clifford Hoyle, Professor H. Raistrick, Dr. J. G. Scadding, Professor W. H. Tytler, Professor G. S. Wilson, and Dr. P. D'Arcy Hart (secretary). The centres at which the work was carried out and the specialists in charge of patients and pathological work were as follows:

Brompton Hospital, London.—Clinician: Dr. J. W. Crofton, Streptomycin Registrar (working under the direction of the honorary staff of Brompton Hospital); Pathologists: Dr. J. W. Clegg, Dr. D. A. Mitchison.

Colindale Hospital (L.C.C.), London.—Clinicians: Dr. J. V. Hurford, Dr. B. J. Douglas Smith, Dr. W. E. Snell; Pathologists (Central Public Health Laboratory): Dr. G. B. Forbes, Dr. H. D. Holt.

Harefield Hospital (M.C.C.), Harefield, Middlesex.—Clinicians: Dr. R. H. Brent, Dr. L. E. Houghton; Pathologist: Dr. E. Nassau.

Bangour Hospital, Bangour, West Lothian.—Clinician: Dr. I. D. Ross; Pathologist: Dr. Isabella Purdie.

Killingbeck Hospital and Sanatorium, Leeds.—Clinicians: Dr. W. Santon Gilmour, Dr. A. M. Reevie; Pathologist: Professor J. W. McLeod.

Northern Hospital (L.C.C.), Winchmore Hill, London.—Clinicians: Dr. F. A. Nash, Dr. R. Shoulman; Pathologists: Dr. J. M. Alston, Dr. A. Mohun.

Sully Hospital, Sully, Glam.—Clinicians: Dr. D. M. E. Thomas, Dr. L. R. West; Pathologist: Professor W. H. Tytler.

The clinicians of the centres met periodically as a working subcommittee under the chairmanship of Dr. Geoffrey Marshall; so also did the pathologists under the chairmanship of Dr. R. Cruickshank. Dr. Marc Daniels, of the Council's scientific staff, was responsible for the clinical co-ordination of the trials, and he also prepared the report for the Committee, with assistance from Dr. D. A. Mitchison on the analysis of laboratory results. For the purpose of final analysis the radiological findings were assessed by a panel composed of Dr. L. G. Blair, Dr. Peter Kerley, and Dr. Geoffrey S. Todd.

“If your experiment needs statistics, then you ought to have done a better experiment”

Ernest Rutherford (1871-1937)

Therein lies the rub….

With the notable exception of psychology, and to a lesser extent economics, in the social sciences experimentation is often not routinely possible

(e.g. we cannot randomise people to ethnic and gender groups, social housing, schools, local authorities etc. etc.)

The social world is complex!

In the non-experimental social sciences we must use more comprehensive statistical methods which might better help us to identify, and then quantify, the multifaceted relationships that characterise contemporary social life

Is there a relationship between fat intake and breast cancer?

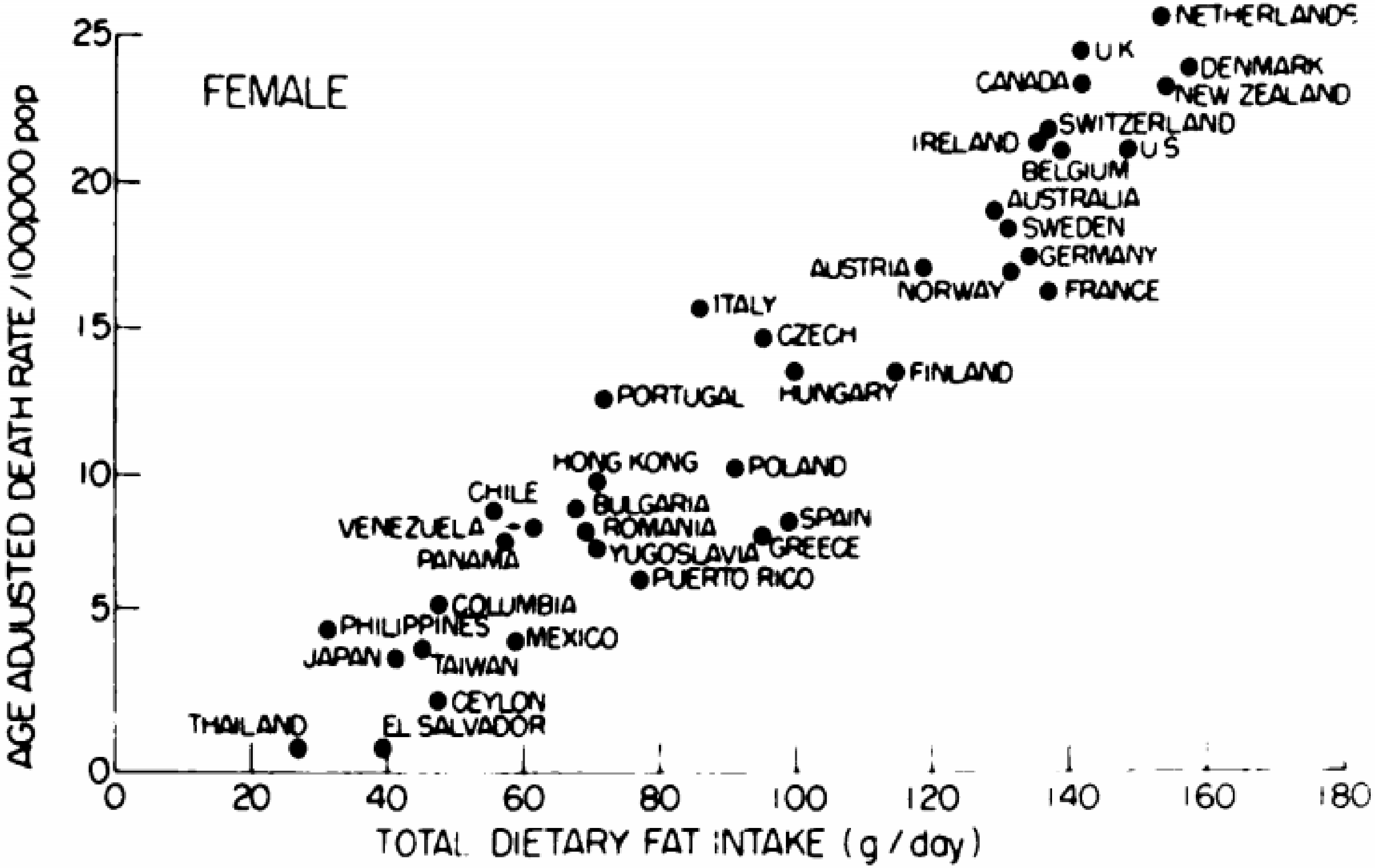


Chart 3. Correlation between per capita consumption of dietary fat and age-adjusted mortality from breast cancer in different countries

Carroll, K. (1975) ‘Experimental Evidence of Dietary Factors and Hormone-dependent Cancers’, Cancer Research, 35, pp.3374-3383.

Is there a relationship between fat intake and breast cancer?

Yes all other things being equal

But all other things are not equal, *or are they*?

Is there a relationship between fat intake and breast cancer?

For example countries with a lot of fat in their diet might also have a lot of sugar in their diet

In richer countries people tend to eat more fat and more sugar

Risks

- Recent use of birth control pills
- Not having children / late childbirth
- Not breastfeeding
- Alcohol use
- Being overweight or obese

Suggested Risks?

- Diet
- Antiperspirants
- Bras
- Pollution
- Tobacco smoke
- Night work

The American Cancer Society website (www.cancer.org) suggests a series of Breast Cancer Risks related to lifestyle choices

Thinking Predictively

Simple question –

but deceptively difficult to answer

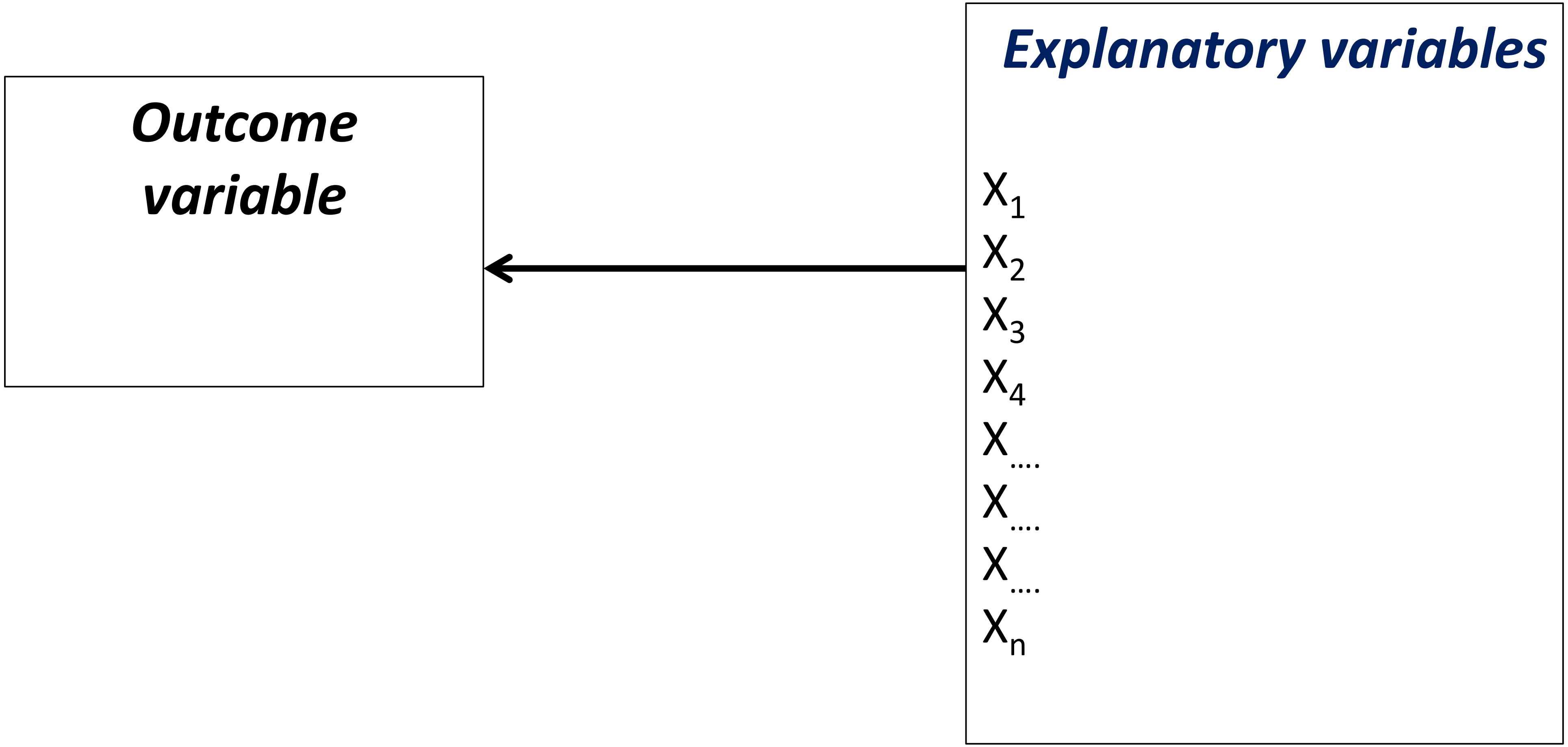
What lies on the causal pathway?

Thinking Predictively

What lies on the causal pathway?

- **Experimental Data** are easy – it is part of the design!
- **Observational Data** require ‘sophisticated multivariate analyses’

We need a statistical model
i.e. a multivariate statistical approach



Outcome variable

Explanatory variables

X_1

X_2

X_3

X_4

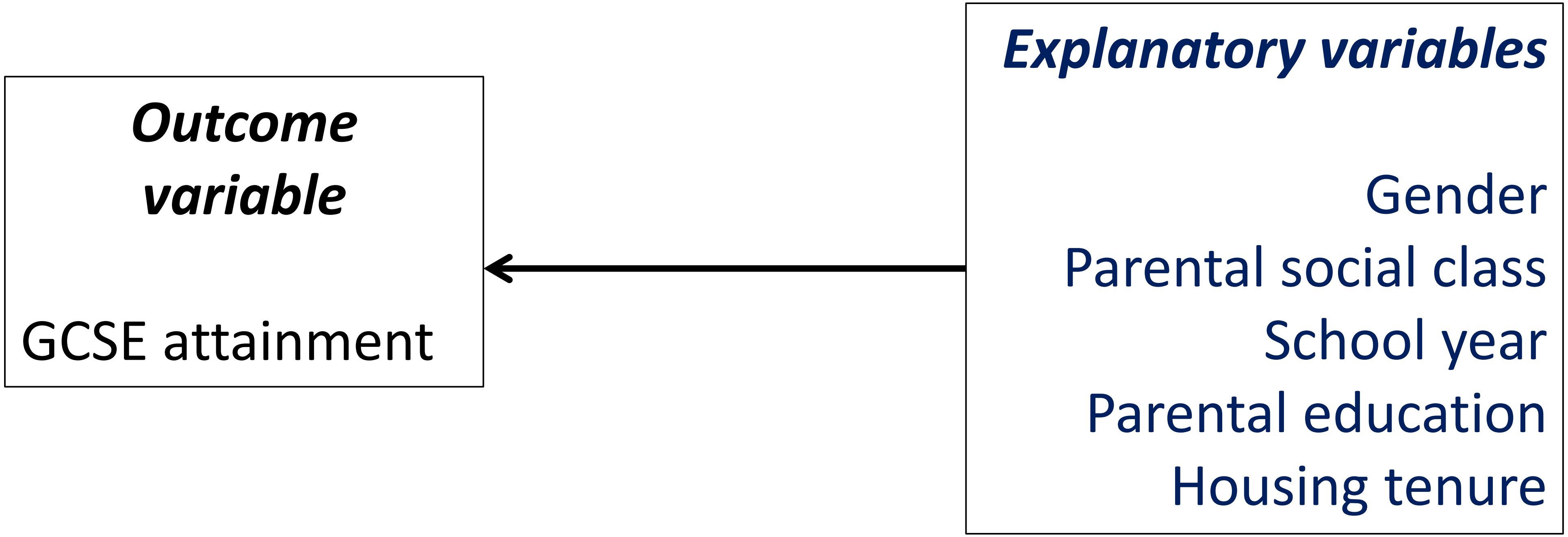
$X_{....}$

$X_{....}$

$X_{....}$

X_n

Examples variables in a real paper



Outcome variable

GCSE attainment

Explanatory variables

Gender

Parental social class

School year

Parental education

Housing tenure

Connelly, Murray and Gayle (2013) *Sociological Research Online*

Sir Francis Galton (1822-1911)

- Darwin’s cousin
- Developed finger printing
- First weather map (Times 1st April 1875)
- Cutting a Round Cake on Scientific Principles (Nature 1906)
- Strawberry Cure for Gout (Nature 1899)
- On Spectacles for Divers
- Beauty Map of Britain (*I found London to rank highest for beauty: Aberdeen lowest, Memoire p.153)*

A Statistical Model - AKA

Simplest statistical model

- A regression model
- Multiple regression
- Linear regression model
- General linear model
- Vanilla regression

A (slightly drunk) statistician once said to me “Vernon, if we didn’t have so many confusing terms we couldn’t charge high consultancy fees”

Writing Down A Simple Statistical Model

$Y_i = \beta_0+\beta_1X_{1i} +\ldots+ \beta_kX_{ki} + \varepsilon_i$

These equations are all Greek to me…

Left Hand Side = Right Hand Side + Error

Left Hand Side (e.g. outcome variable) ….

Y

Left Hand Side (possibly add a subscript)

$Y_i =$

Constant

$$Y_i = \beta_0$$

First X variable

$$Y_i = \beta_0 + \beta_1 X_{1i}$$

More X variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Right Hand Side

$$= \beta_0+\beta_1 X_{1i} +\ldots+ \beta_k X_{ki}$$

Error term

$$Y_i = \beta_0+\beta_1X_{1i} +\ldots+ \beta_kX_{ki} \boldsymbol{+ \varepsilon_i}$$

Left Hand Side = Right Hand Side + Error

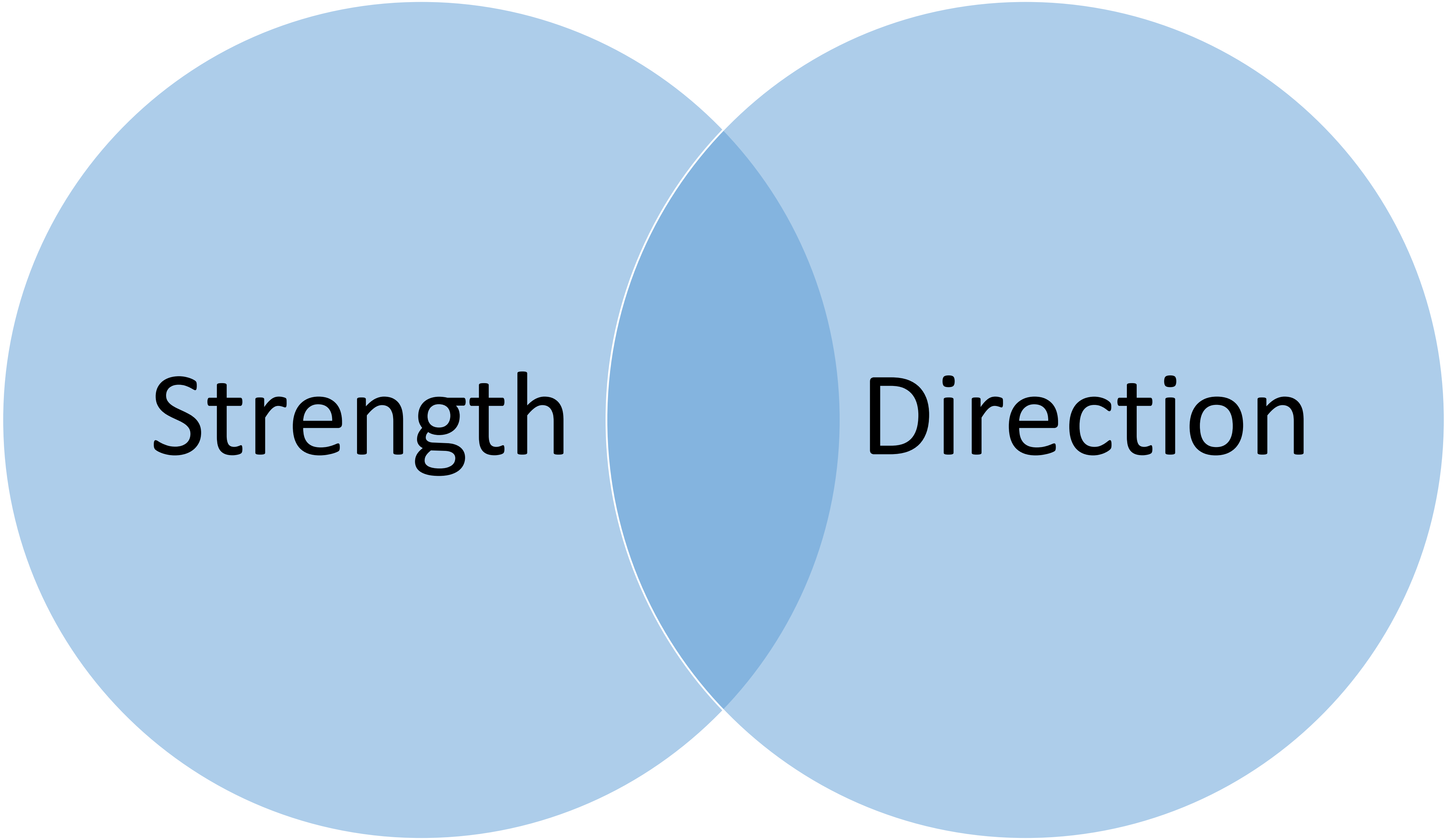
$$Y_i = \beta_0+\beta_1 X_{1i} +\ldots+ \beta_k X_{ki} + \varepsilon_i$$

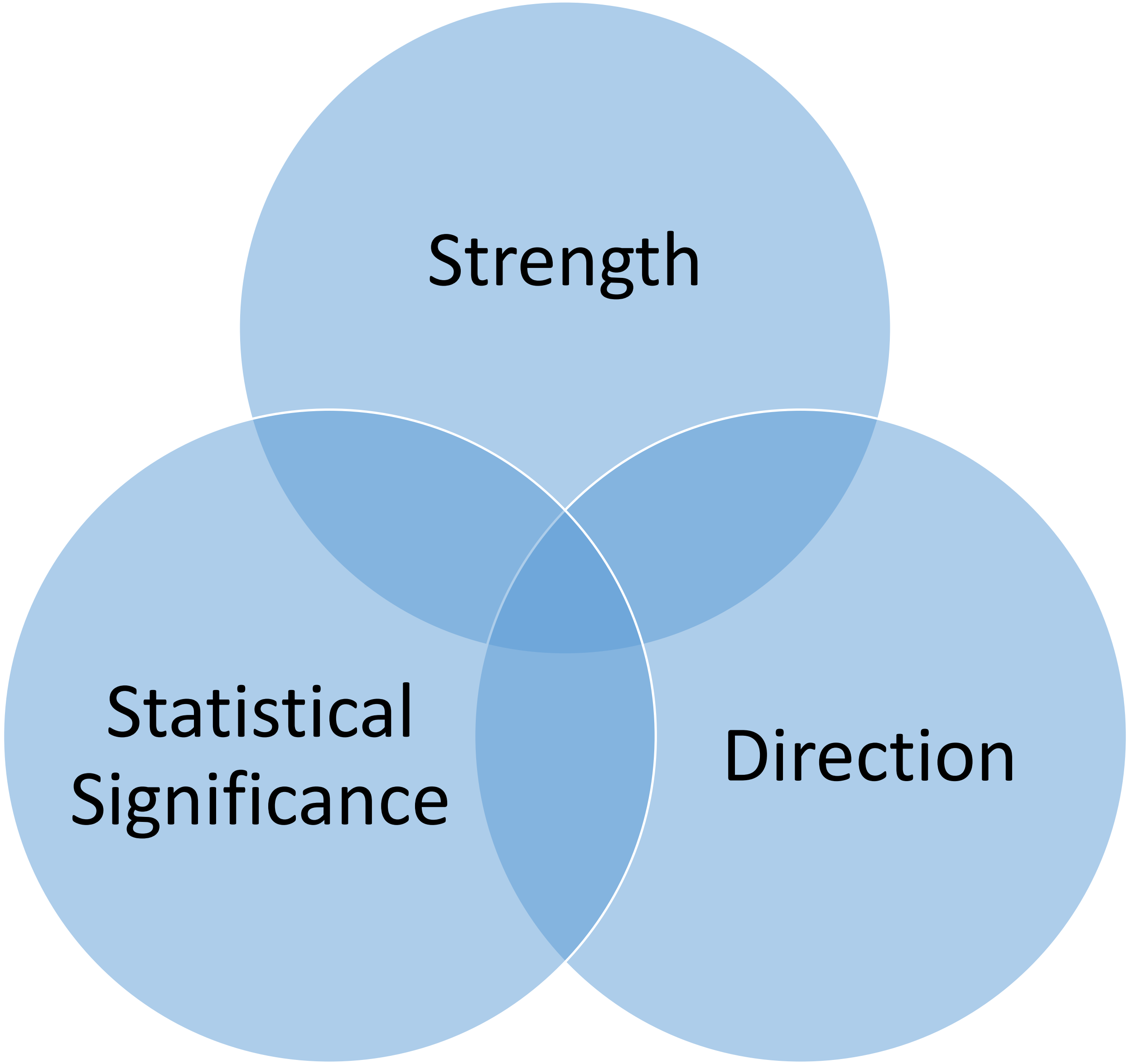
Decoding a more exotic model …

$$Y_{it} = \beta_0 + \lambda_i \ + \beta_1 X_{1it} + \ldots + \beta_k X_{kit} + \varepsilon_{it}$$

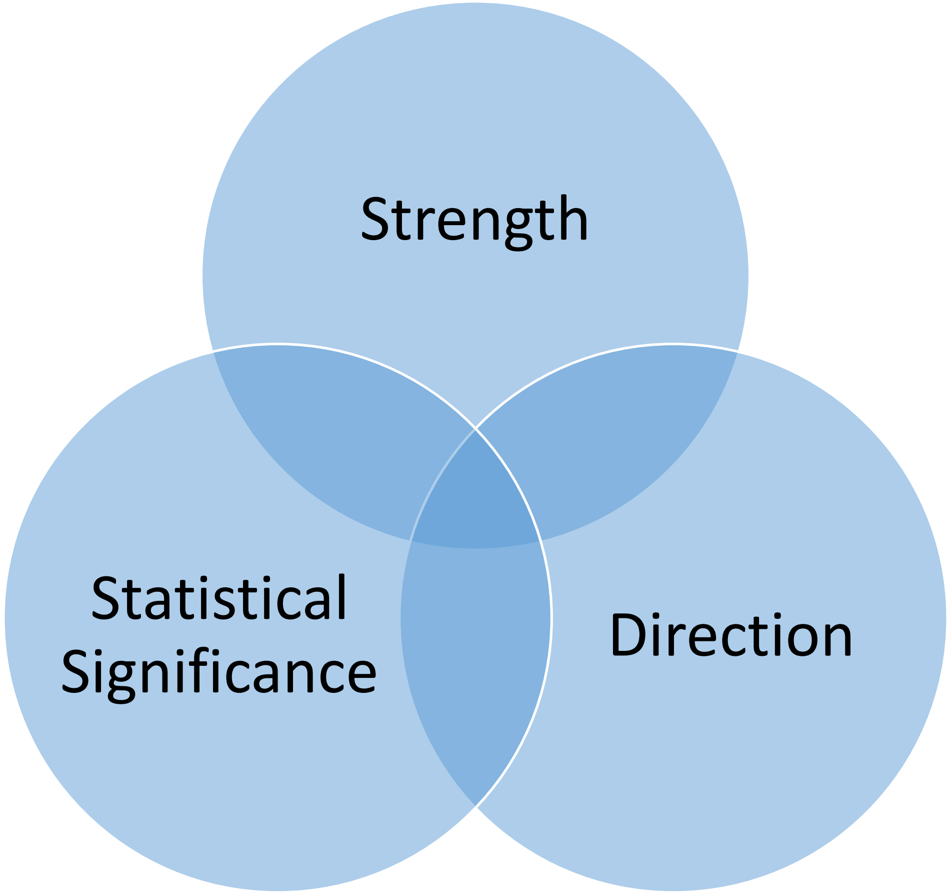
Statistical Models in a Nutshell

Strength


Strength
Direction


Strength
Statistical Significance
Direction

Is what we’ve uncovered important?



Strength

Statistical Significance

Direction

Four Important Components of a Statistical Model

1. Beta (β) size for each e**X**planatory variable
2. Beta (β) sign for each e**X**planatory variable
3. p values for each e**X**planatory variable
4. Goodness of fit - R^2 , Adj. R^2, BIC, etc.

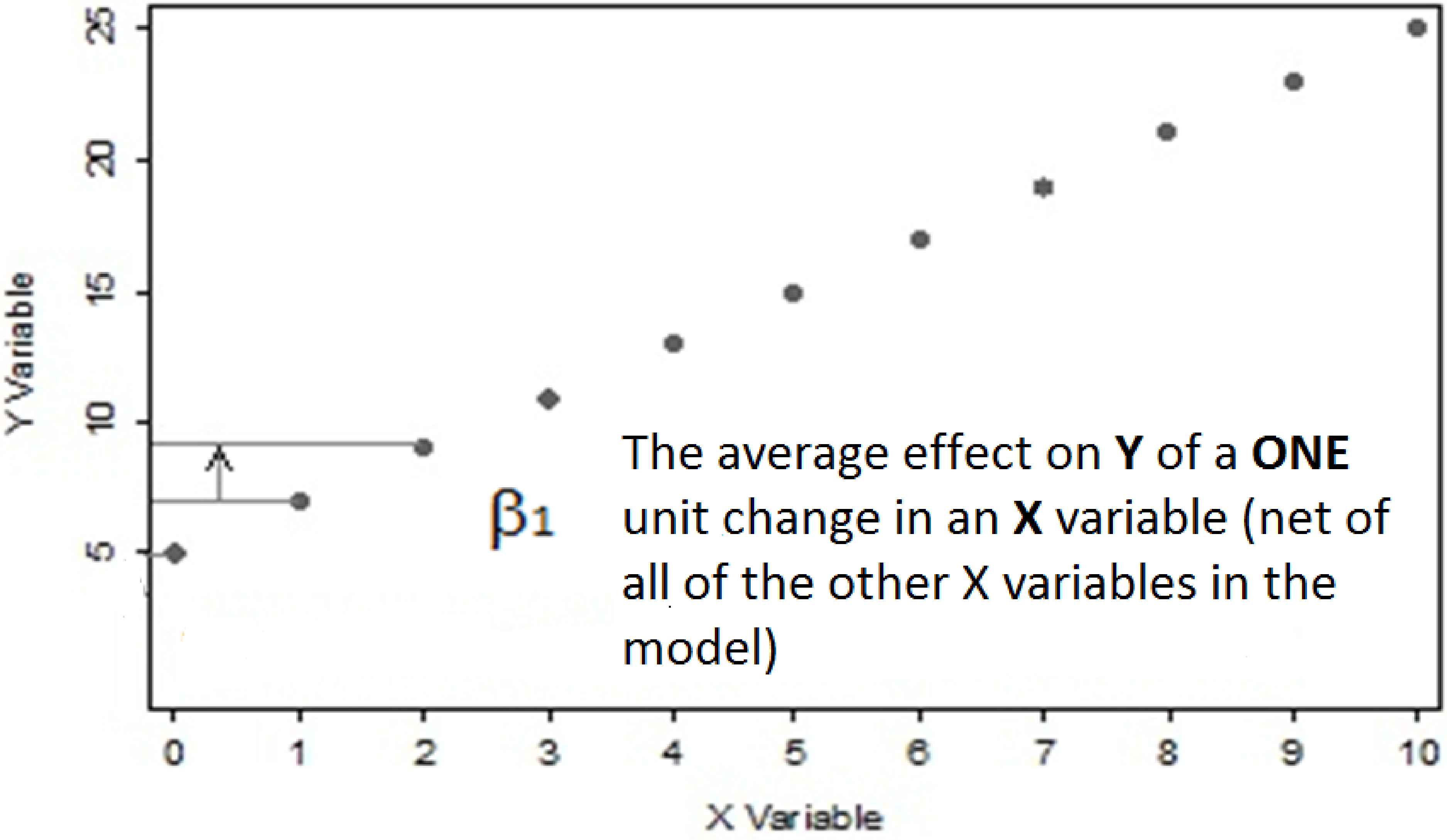
Component # 1

The size of Beta (β)

when Beta (β) is LARGE a one unit change has a BIG effect on Y (net of all the other e**X**planatory variables)

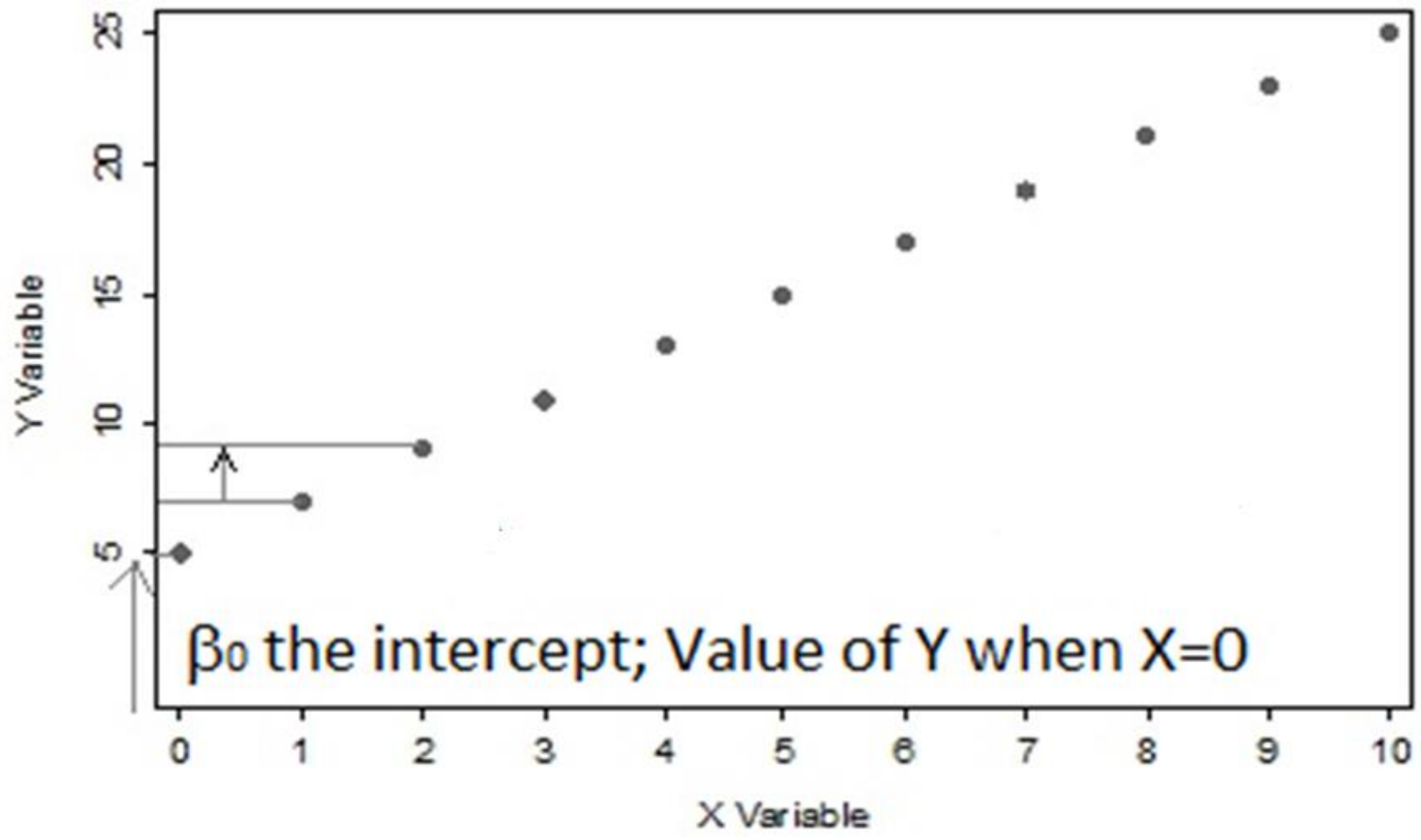
(In some example (β) is standardized so all e**X**planatory variables are on the same scale)

Component # 1 *continued*



β_1 The average effect on **Y** of a **ONE** unit change in an **X** variable (net of all of the other X variables in the model)

The Constant is Beta Zero (β_0)



β_0 the intercept; Value of Y when X=0

Component # 2

- Beta (β) – one for each e**X**planatory variable
- If it is positive (+) increasing X has a positive effect on Y (net of all the other variables)
- If it is negative (-) increasing X has a negative effect on Y (net of all the other variables)

Component # 3

- **p value** for EACH variable in the model

- **p value** tells us if the variable is significant, **NET** of all of the **OTHER** variables in the model

  (The old text books used to say *Ceteris paribus* or 'all other things being equal')

- Debates rage about interpreting the **p value** (i.e. is it less than .05 etc)

Component # 4

- R^2 the Coefficient of Multiple Determination

  Takes on values between 0 and 1

- The proportion of variability in **Y** that is explained by **ALL** of the E**X**PLANATORY variables in the model

- In practice a R^2 value of .25 could be quite high in many studies

Component # 4

- Adjusted R^2

- BIC Bayesian Information Criterion

- AIK Akaike’s Information Criterion

Four Important Components of a Statistical Model

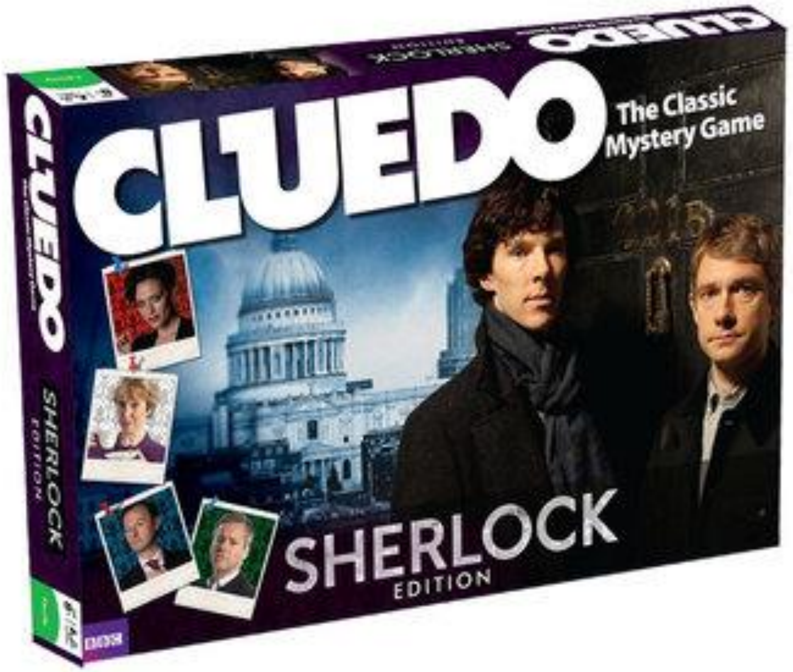
1. Beta (β) size for each e**X**planatory variable
2. Beta (β) sign for each e**X**planatory variable
3. p values for each e**X**planatory variable
4. Goodness of fit - R^2 , Adj. R^2, BIC, etc.

Model Building

- John Tukey – Exploratory Data Analysis

- X variables should have the means, the motive and the opportunity to commit the crime of changing the Y variable –

  Robert Luskin, U. of Texas



Which Model…

Generally…..

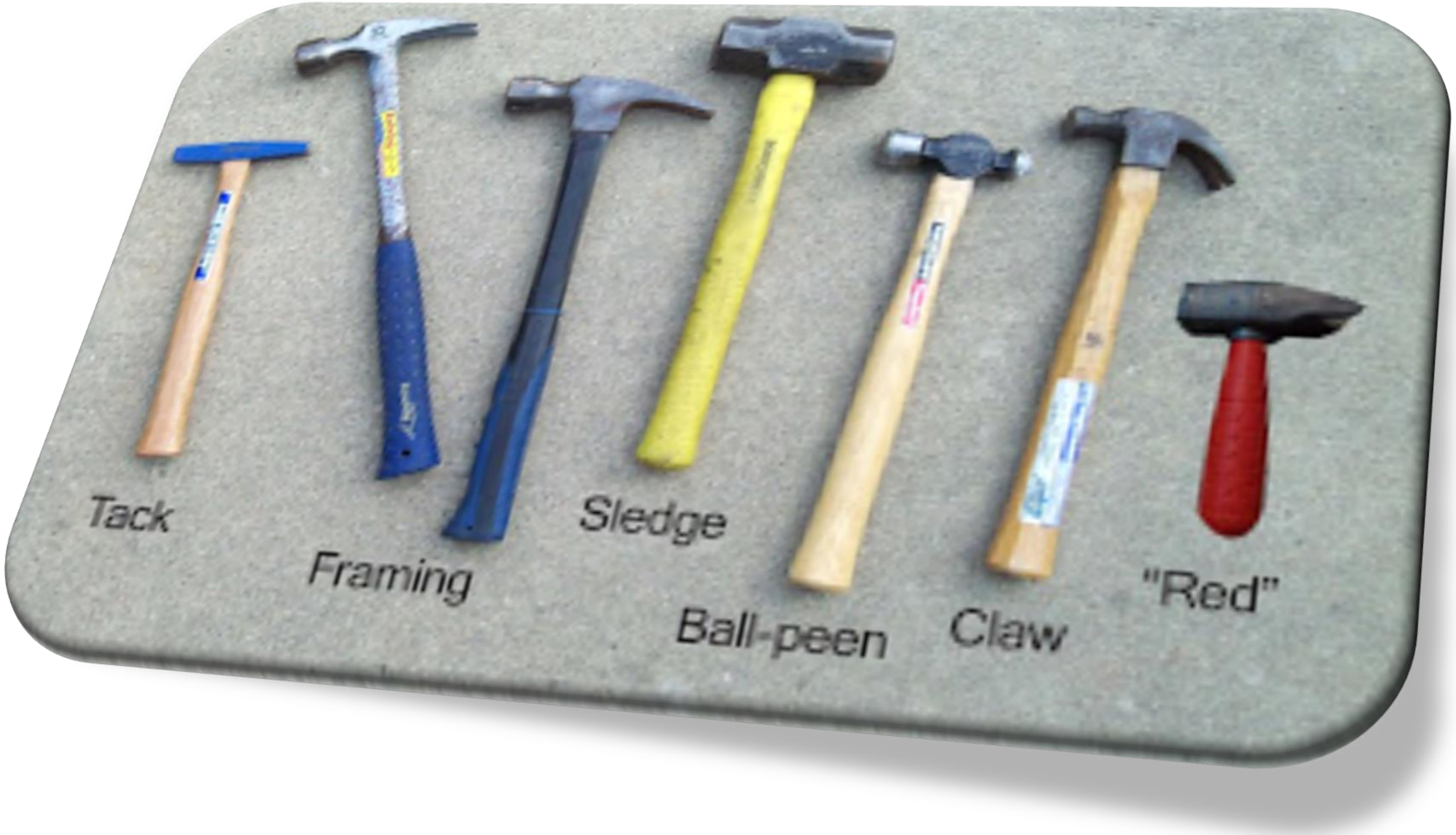
Depends on the outcomes (or outcomes) you are modelling?

Depends on the structure of your data?

Depends on the underlying process that has generated the data?

RENT-A-RELIC
NOTARY
4444

Which tool?



Generalized Linear Latent and Mixed Models

Generalized Linear Mixed Models

Generalized Linear Model

MODEL	OUTCOME	DATA STRUCTURE	PROCESS GENERATING THE DATA
Linear Regression	Metric	Tabular – wide format	Standard cross-section (e.g. one off)
Logistic Regression	Binary (0,1)	Tabular – wide format	Standard cross-section (e.g. one off)
Poisson Regression	Count (e.g. beach visits)	Tabular – wide format	Standard cross-section (e.g. one off)
Panel Regression	Various	Tabular – long format	Repeated contacts
Duration Regression	Time or 'Hazard'	Tabular – wide or long format	Time to event

Workshop

R

- Growing in popularity (e.g. data science, statistics, science etc.)
- Popular with statisticians
- Free (open source)
- Difficult to learn
- Development and support is not commercial
- Help resources are under-developed

Programme

Mix of talks and self-directed practical activities.

Predictive Analytics challenge (“Hackathon”).

Tutor and peer support.

Use of a variety of data sets, especially messy administrative records.

Top tips

1. Ask plenty of questions.
2. Take your time.
3. Complete as many of the tasks and exercises, and answer as many of the questions as you can.
4. Annotate your work.
5. Be positive.

Estimating Work Time…

Good Luck

Our aim is to equip you, as rapidly and painlessly as possible, with a proficiency in predictive analytics using R.

We think it is an ambitious yet achievable goal.

Them: “Are you any good at predictive analytics/statistical modelling?”

You: “Yes, yes I am.”