An Introduction to Statistical Modelling for Qualitative Researchers

Professor Vernon Gayle

vernon.gayle@ed.ac.uk
@Profbigvern
github.com/vernongayle

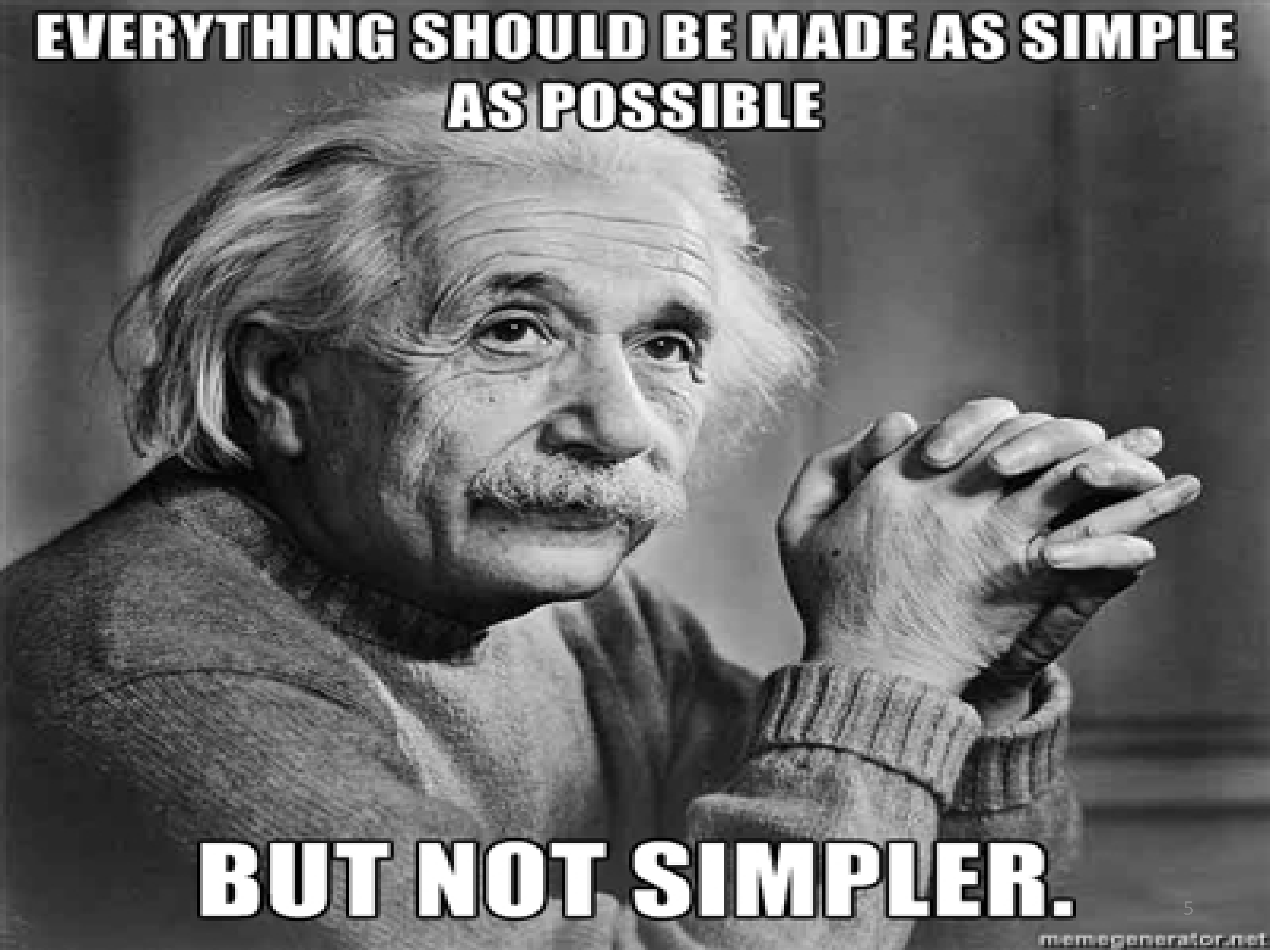
AQMEN

THE UNIVERSITY
of EDINBURGH

Part 1 The Preamble

facebook
December 2010

A semester course in a day!


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

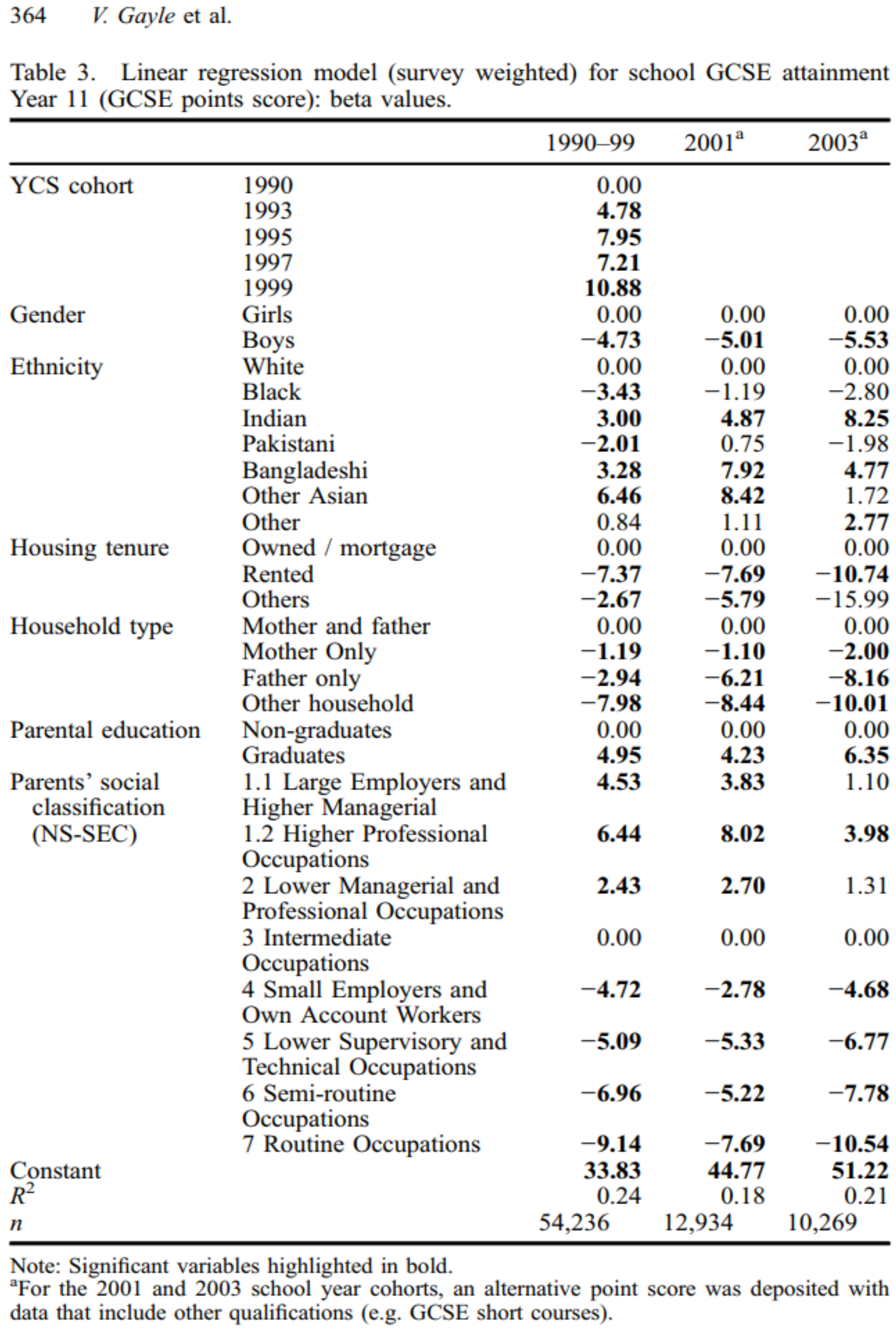


Table 3. Linear regression model (survey weighted) for school GCSE attainment Year 11 (GCSE points score): beta values.

		1990–99	2001[a]	2003[a]
YCS cohort	1990	0.00		
	1993	**4.78**		
	1995	**7.95**		
	1997	**7.21**		
	1999	**10.88**		
Gender	Girls	0.00	0.00	0.00
	Boys	**−4.73**	**−5.01**	**−5.53**
Ethnicity	White	0.00	0.00	0.00
	Black	**−3.43**	−1.19	−2.80
	Indian	**3.00**	**4.87**	**8.25**
	Pakistani	**−2.01**	0.75	−1.98
	Bangladeshi	**3.28**	**7.92**	**4.77**
	Other Asian	**6.46**	**8.42**	1.72
	Other	0.84	1.11	**2.77**
Housing tenure	Owned / mortgage	0.00	0.00	0.00
	Rented	**−7.37**	**−7.69**	**−10.74**
	Others	**−2.67**	**−5.79**	−15.99
Household type	Mother and father	0.00	0.00	0.00
	Mother Only	**−1.19**	**−1.10**	**−2.00**
	Father only	**−2.94**	**−6.21**	**−8.16**
	Other household	**−7.98**	**−8.44**	**−10.01**
Parental education	Non-graduates	0.00	0.00	0.00
	Graduates	**4.95**	**4.23**	**6.35**
Parents' social classification (NS-SEC)	1.1 Large Employers and Higher Managerial	**4.53**	**3.83**	1.10
	1.2 Higher Professional Occupations	**6.44**	**8.02**	**3.98**
	2 Lower Managerial and Professional Occupations	**2.43**	**2.70**	1.31
	3 Intermediate Occupations	0.00	0.00	0.00
	4 Small Employers and Own Account Workers	**−4.72**	**−2.78**	**−4.68**
	5 Lower Supervisory and Technical Occupations	**−5.09**	**−5.33**	**−6.77**
	6 Semi-routine Occupations	**−6.96**	**−5.22**	**−7.78**
	7 Routine Occupations	**−9.14**	**−7.69**	**−10.54**
Constant		**33.83**	**44.77**	**51.22**
R^2		0.24	0.18	0.21
n		54,236	12,934	10,269

Note: Significant variables highlighted in bold.

[a] For the 2001 and 2003 school year cohorts, an alternative point score was deposited with data that include other qualifications (e.g. GCSE short courses).

Today there will be…

- No discussion of the limitations of quantitative methods in social science
- No discussion of the limitations of social surveys
- No discussion of data quality
- General data examples
    - Participants from various social sciences
- Plenty of anecdotes (stop me if there are too many)

My aims

- Convey some of my enthusiasm for the topic
- Engage (and possibly even entertain) participants
- Alleviate anxiety a little
- Encourage people to ask questions
- Leaving with a bit more knowledge (end skipping)
- Possibly motivate people to do more in future

The speed of presentations - tell me if it is too fast or too slow!!!

The Take-Home Messages

However, why it is not called the ‘take back to the office message’ is beyond me!

My advice is don’t take a message home take a bottle of wine instead!

Is there a relationship between fat intake and breast cancer?

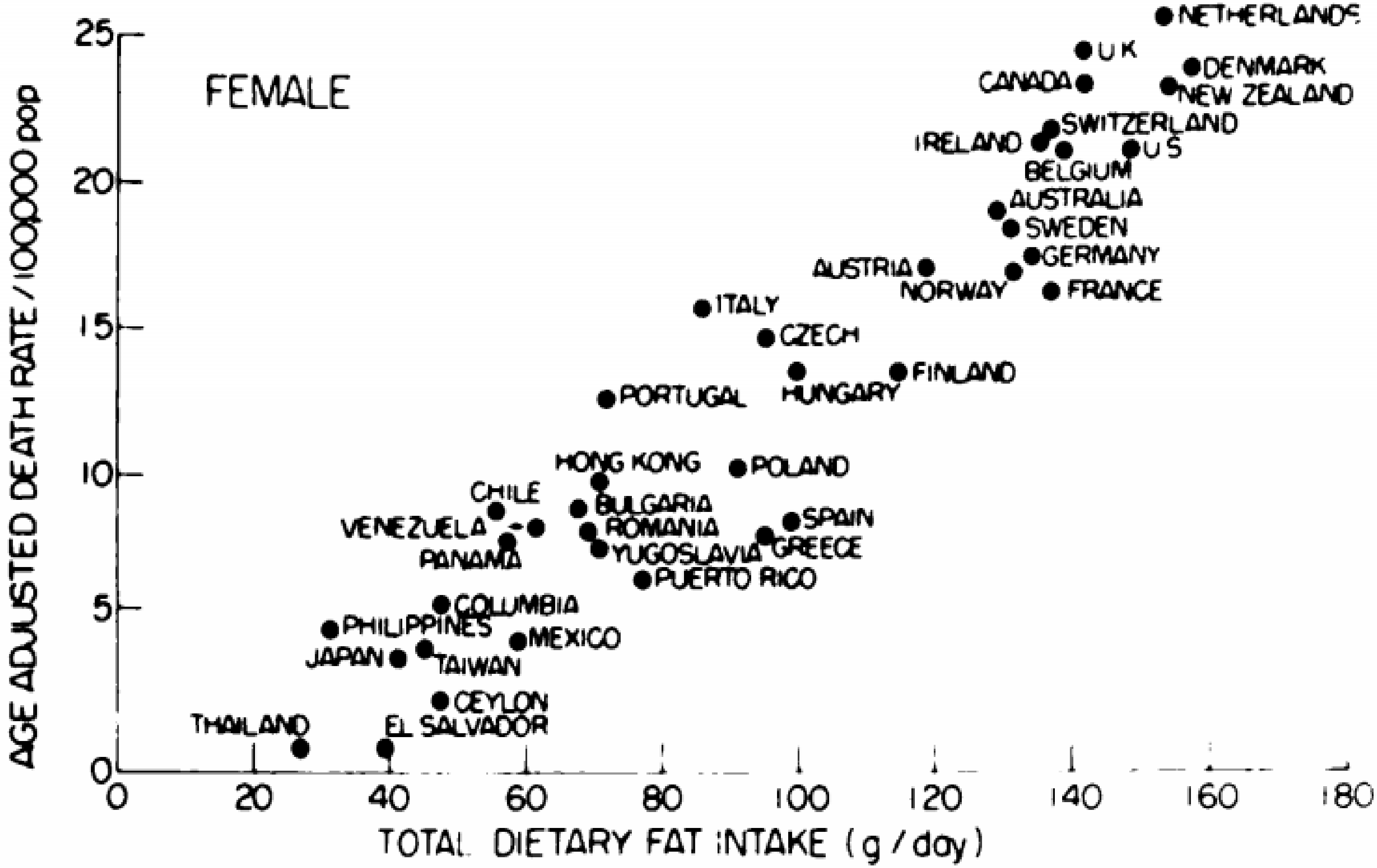


Chart 3. Correlation between per capita consumption of dietary fat and age-adjusted mortality from breast cancer in different countries

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

The American Cancer Society

Suggests a series of Breast Cancer(www.cancer.org)

Risks related to lifestyle choices

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

They also suggest a series of risk factors that are not certain

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

Is there a relationship between fat intake and breast cancer?

Do we suspect that in different countries not everything is equal?

Are we still convinced that eating fat is related to breast cancer?

Do we need more comprehensive statistical analyses?

- i.e. statistical models

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

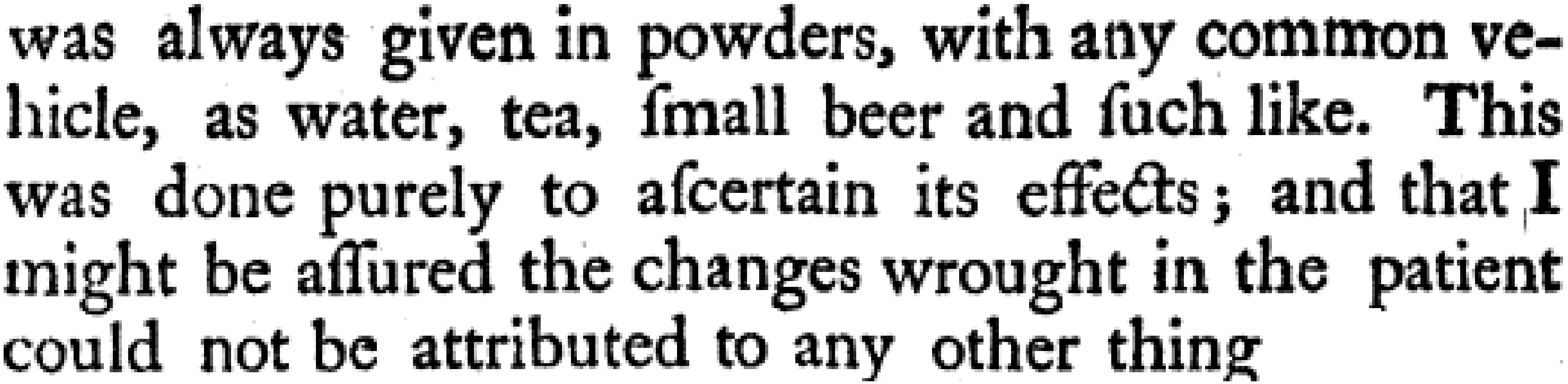
Twelve healthy active subjects (6 male, 6 female) performed 60 min of exercise (60% VO_{2max}) on 3 occasions after supplementing with L-Carnitine L-tartrate (LCLT) or placebo. Each subject received a chronic dose, an acute dose, and placebo in a randomized, double-blind crossover design. Dietary intake 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,

Chipping-Norton,
Oxfordſhire,
April 25, 1763.

your Lordſhip's moſt obedient
humble Servant

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 Take-Home Message #1

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

Part 2 Basic Concepts in Statistical Data Analysis

Basic Concepts (probably revision)

- Variables
    - measures of social science concepts
- Cases
    - Distinctive entities
    - People, firms, farms, hospitals, schools, local authorities, regions, nation states, animals

- Population N
  - UK (Decennial) Census
  - All the police officers in Scotland

- Sample n
  - 10% of all of the police officers in Scotland

- Census (whole population)
- Social Survey (usually a sample)
- Administrative source might cover all or part of population

Basic Concepts

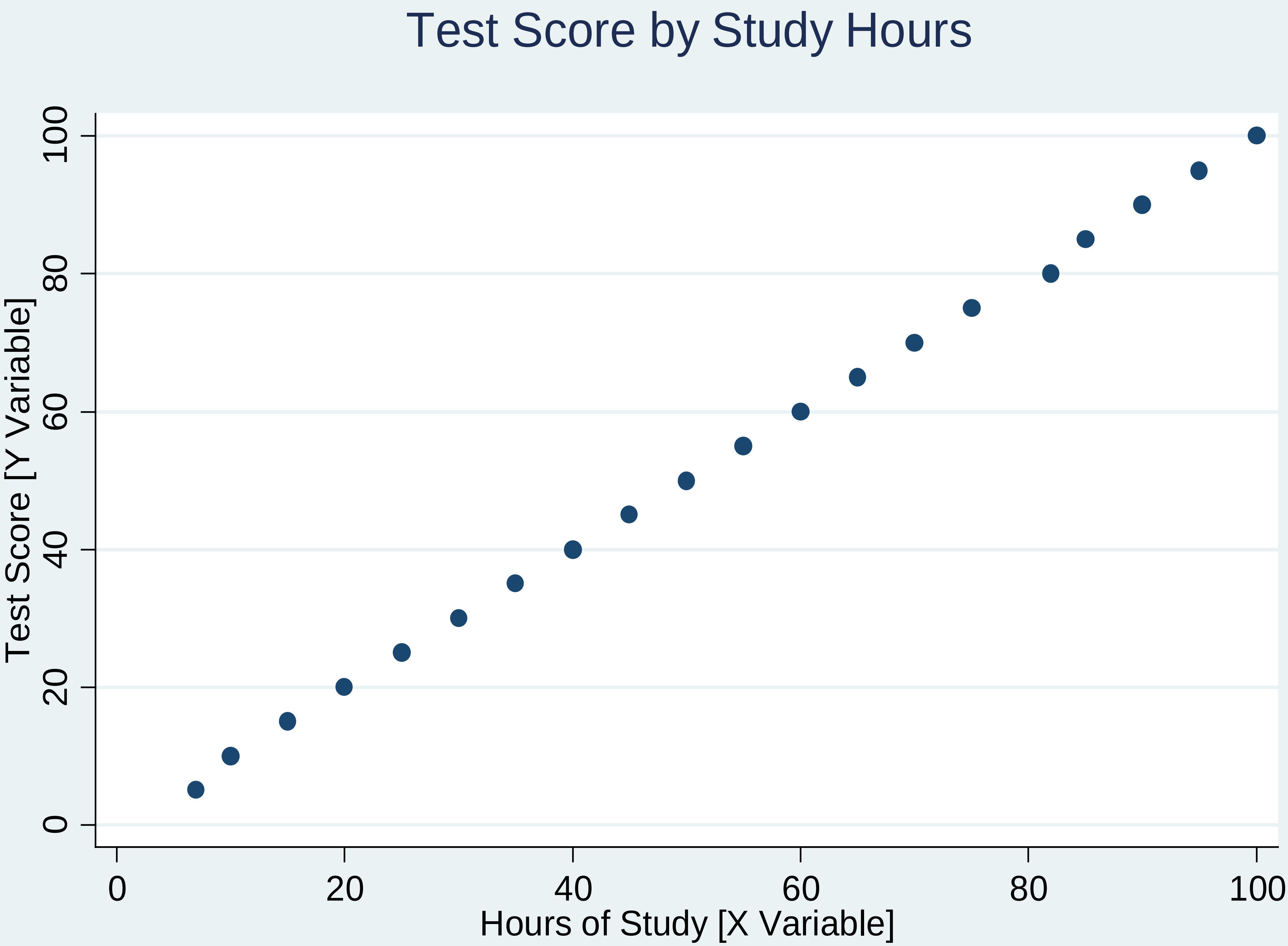
- Outcome variables
  - Y variables

- Educational test score
- Life expectancy (years)
- Number of criminal convictions
- Numerous health outcomes
- Subjective wellbeing (SWB) measures

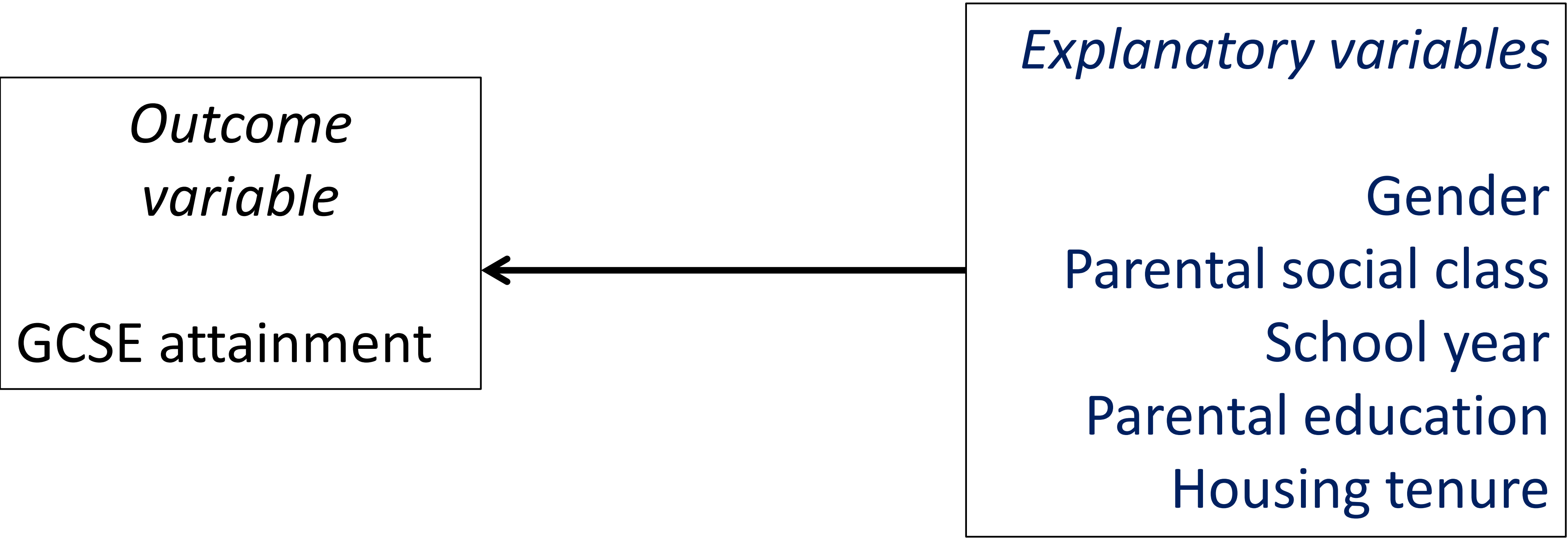
Basic Concepts

- Explanatory variables
  - X variables
  - These variables explain outcome variables

- Hours of study
- Gender
- Ethnicity
- Socioeconomic classifications
- Age
- Housing tenure (type)


Test Score by Study Hours
Test Score [Y Variable]
0
20
40
60
80
100
0
20
40
60
80
100
Hours of Study [X Variable]

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*

- **Univariate** – a single variable

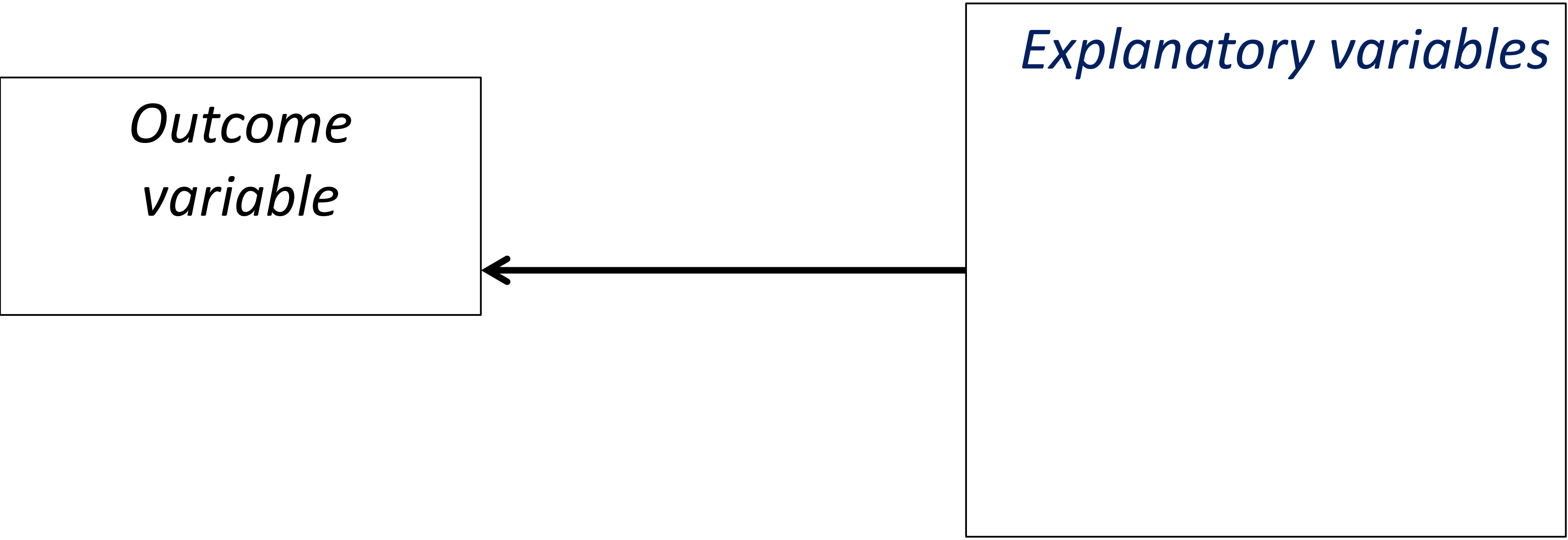
- **Bivariate** – two variables
  - One outcome variable (Y) and one explanatory variable (X)

- **Multivariate** – three (or many more) variables
  - One outcome variable (Y) and many explanatory variables (X)
  - This is the ‘cheddar’ (see Urban Dictionary)

(More advanced multivariate analyses have multiple outcomes too)

Activity 1

1. Identify an outcome variable in your research area
2. Think of a handful of plausible explanatory variables
3. Think of an implausible explanatory variable
4. Is there a dataset (or data source) which includes these variables?



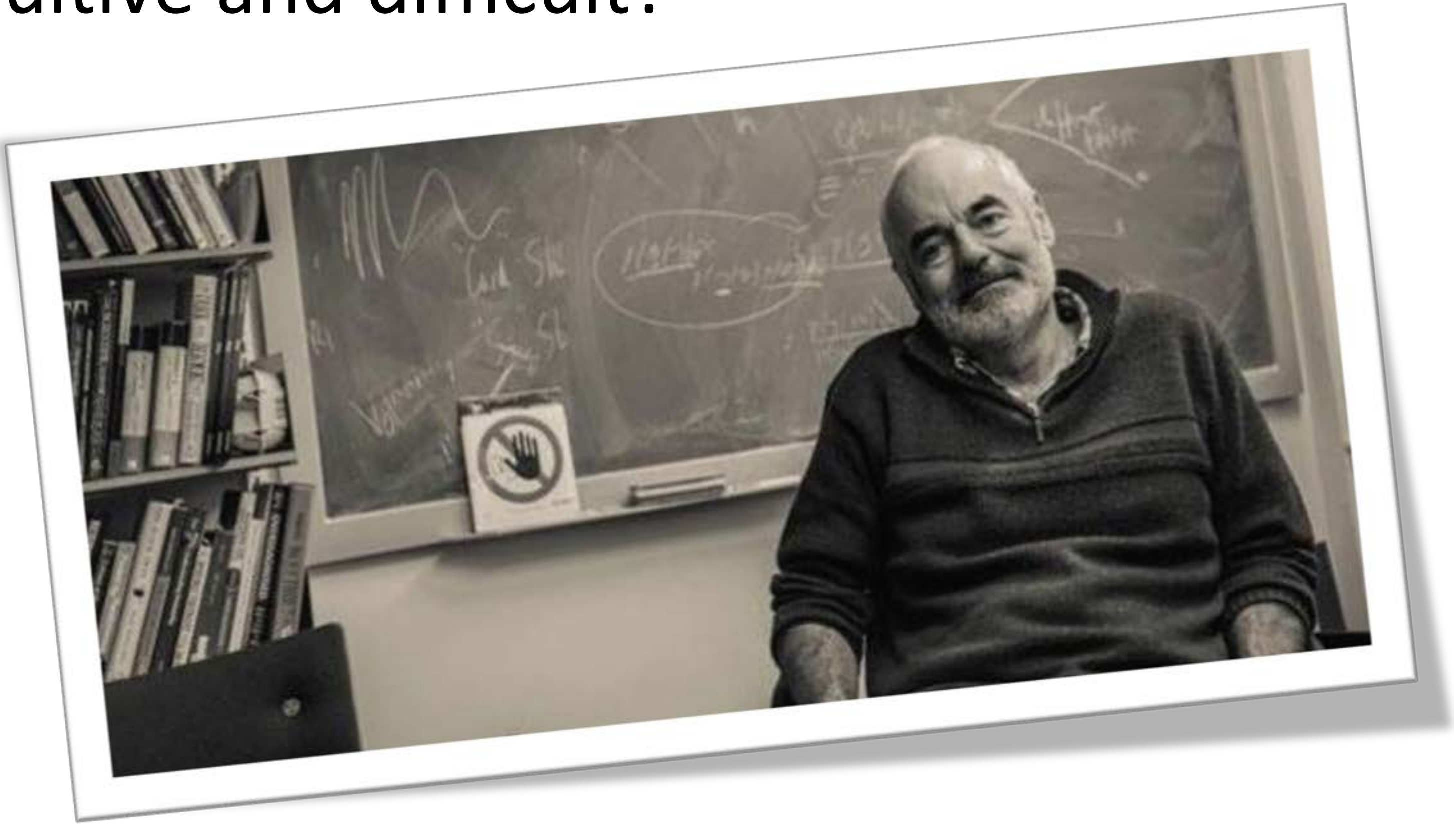
Outcome variable

Explanatory variables

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

Part 3 Probability

Why do so many people find probability theory so unintuitive and difficult?

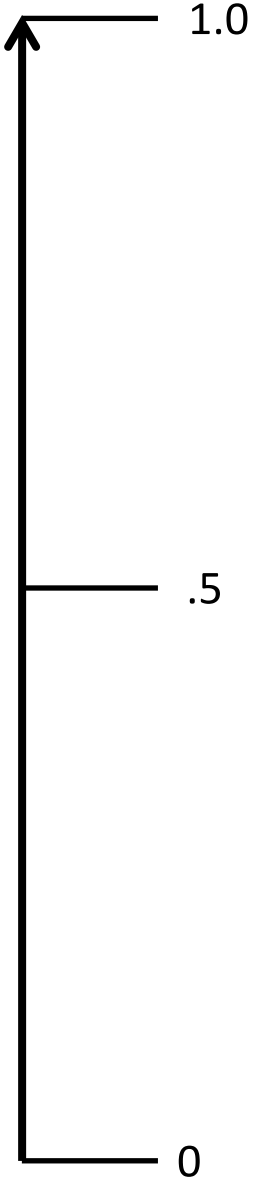


After years of careful study, I have finally found it's because probability is unintuitive and difficult.

http://www.wired.co.uk/magazine/archive/2011/09/ideas-bank/david-spiegelhalter-probability-is-likely-to-confuse-people

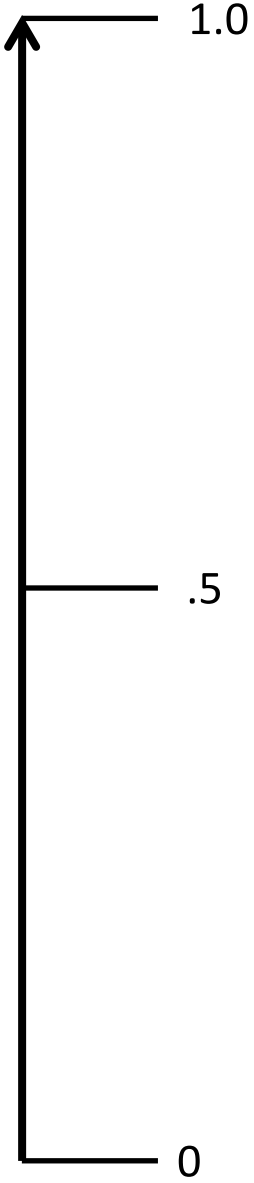
“It is clear talking to people that I am not the only one who finds probability difficult. I find probability problems very difficult indeed. I have to sit and think carefully. Usually because there are two or three different ways of doing it, and different ways of approaching the problem. Which is nice, but it is quite difficult.”

Lecture to South African Statisticians



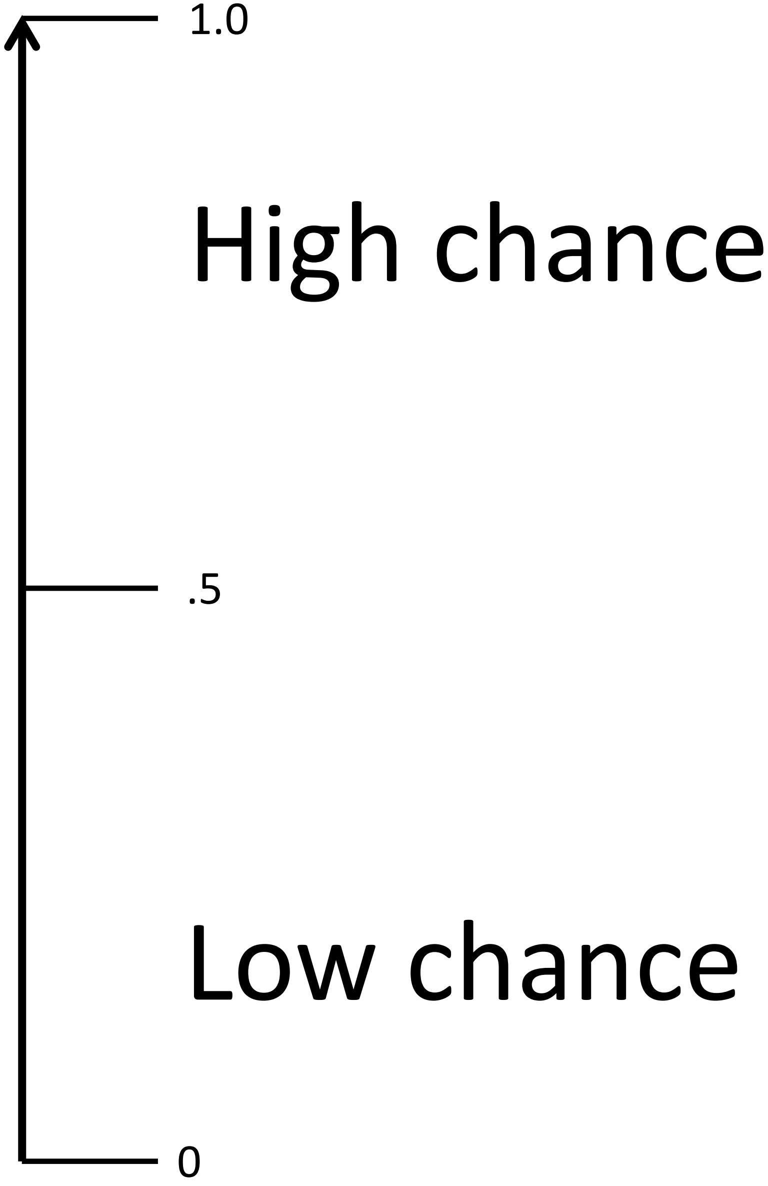
Probabilities take on values between 0 and 1

Denoted a “p”



Event will definitely occur

Event will definitely **not** occur



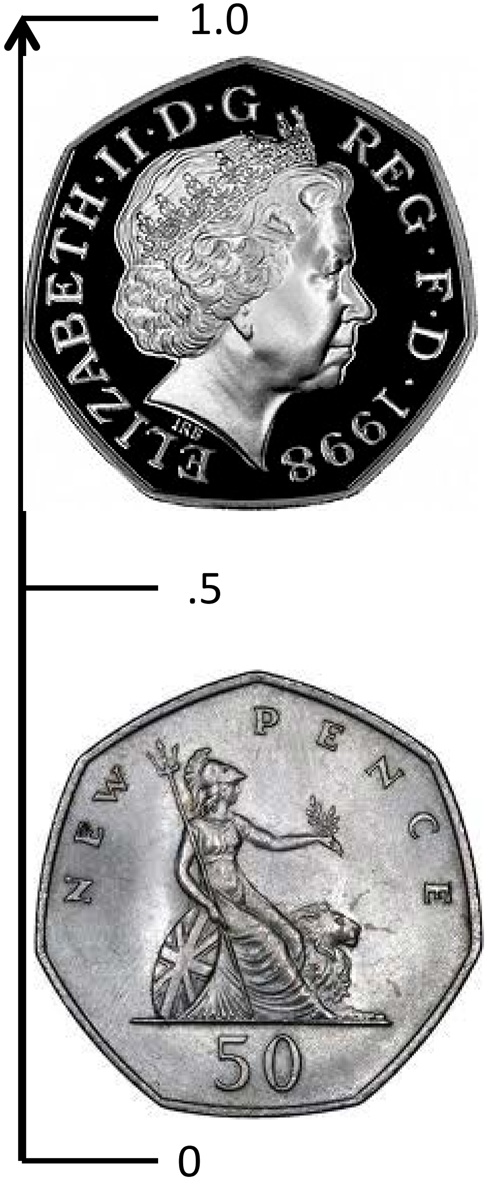
1.0

High chance

.5

Low chance

0



p=.50 even chance

Activity 2

1. Drawing an ace from a single standard pack?

2. Rolling a three with a (fair) single die?

3. Probability of tossing two heads in a row with a 50p coin?

Guess the Probability?

1. Drawing an ace from a single standard pack

   p=.08 (4/52)

2. Rolling a three with a (fair) single die?

   p=.17 (1/6)

3. Probability of tossing two heads in a row with a 50p coin?

   p=.25 (HH; HT; TT; TH = ¼ or ½ * ½)

Probability Distribution of Outcomes Single Roll of a Pair of (fair) Dice

Score	Probability	(p)
2	1/36	.02778
3	2/36	.05556
4	3/36	.08333
5	4/36	.11111
6	5/36	.13889
7	6/36	.16667
8	5/36	.13889
9	4/36	.11111
10	3/36	.08333
11	2/36	.05556
12	1/36	.02778
	36/36	**1.00000**

Probability Distribution of Outcomes
Single Roll of a Pair of (fair) Dice

Score	Probability	(p)
2	1/36	.02778
3	2/36	.05556
4	3/36	.08333
5	4/36	.11111
6	5/36	.13889
7	6/36	.16667
8	5/36	.13889
9	4/36	.11111
10	3/36	.08333
11	2/36	.05556
12	1/36	.02778
	36/36	**1.00000**

Gerolamo Cardano (1501 – 1576) might have been the first person to work this out formally

Scotland v Brazil

If the game ends

Scotland 2 Brazil 2

What is the probability that it was 0 - 0 at half time?

Probability =

Outcome / Total Number of Outcomes

Scotland	Brazil
0	0
0	1
0	2
1	0
1	1
1	2
2	0
2	1
2	2

Probability = Outcome /Total Number of Outcomes

1 / 9 chance it was 0 - 0 at half time

(if all scores are equally likely)

P values

	(Ten)	Hundred				
p=.	9	9				
p=.	2	5				
p=.	1	0				

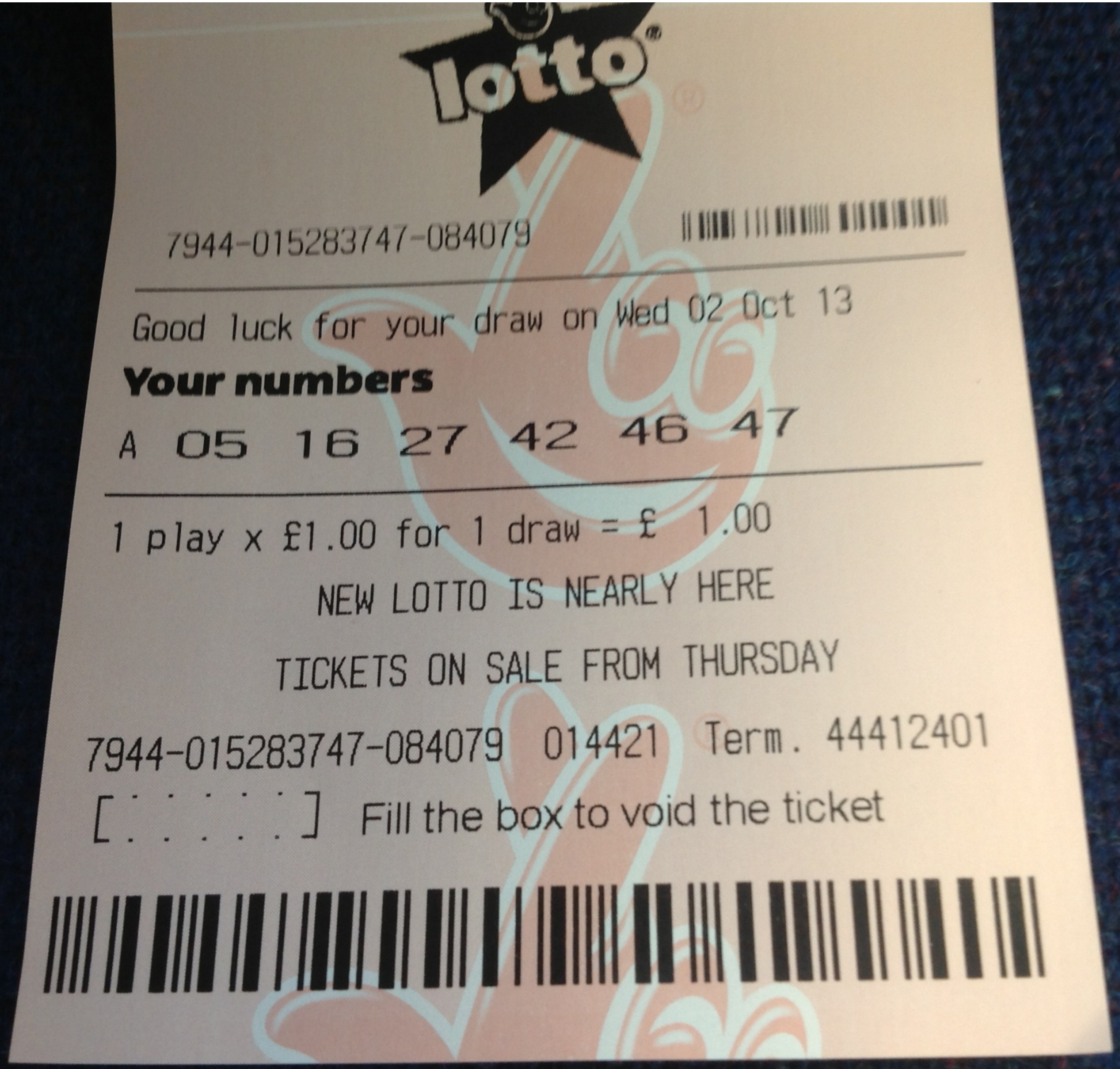
P values

	(Ten)	Hundred	Thousand			
p=.	0	5				
p=.	0	1				
p=.	0	0	1			

P values

	(Ten)	Hundred	Thousand	Ten Thousand	Hundred Thousand	Million
p=.	0	0	0	1		
p=.	0	0	0	0	1	
p=.	0	0	0	0	0	1

1 in a million - 10p tossed 20 times all coming out heads



lotto

7944-015283747-084079

Good luck for your draw on Wed 02 Oct 13

Your numbers

A 05 16 27 42 46 47

1 play x £1.00 for 1 draw = £ 1.00

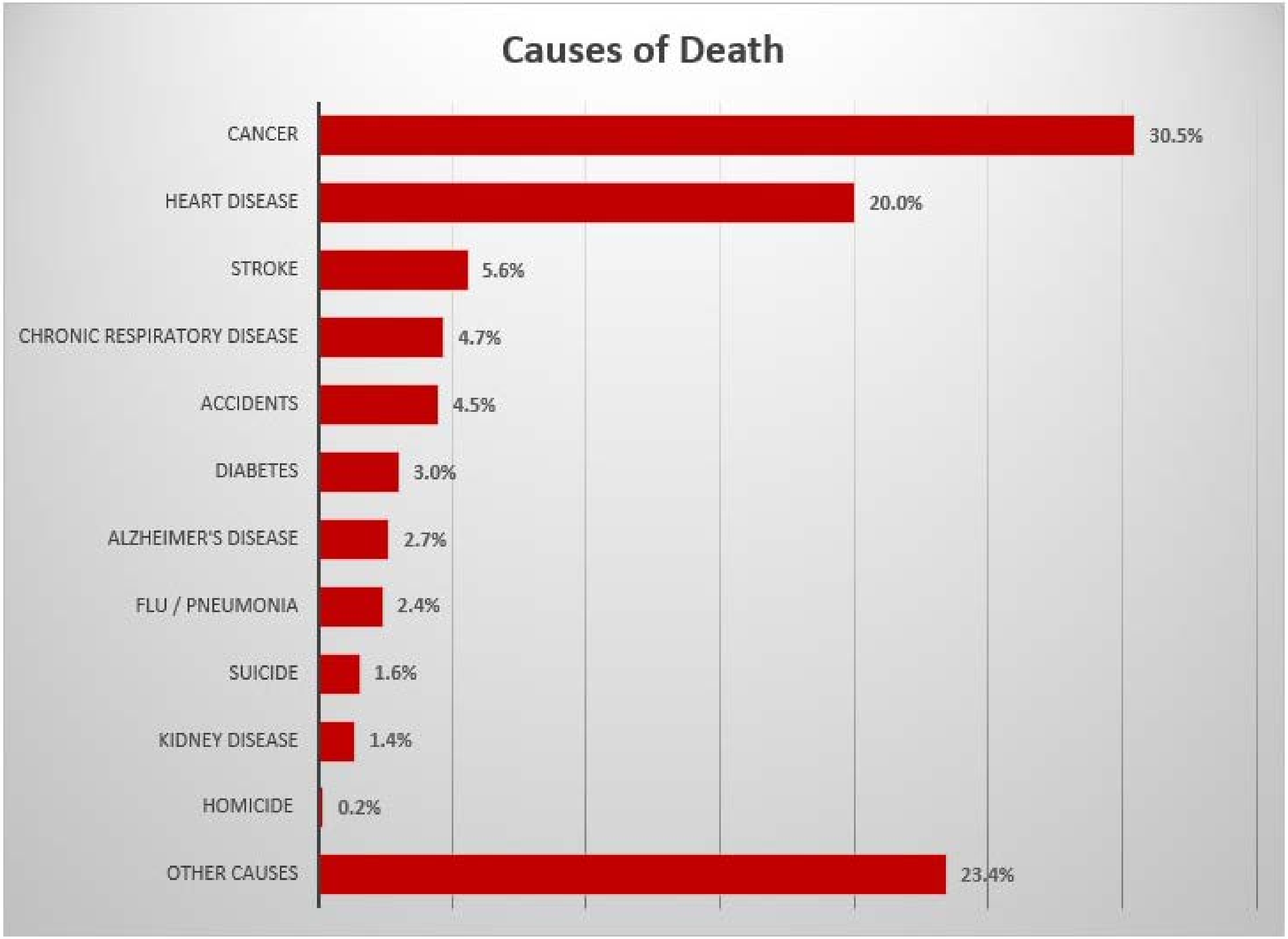
NEW LOTTO IS NEARLY HERE

TICKETS ON SALE FROM THURSDAY

7944-015283747-084079 014421 Term. 44412401

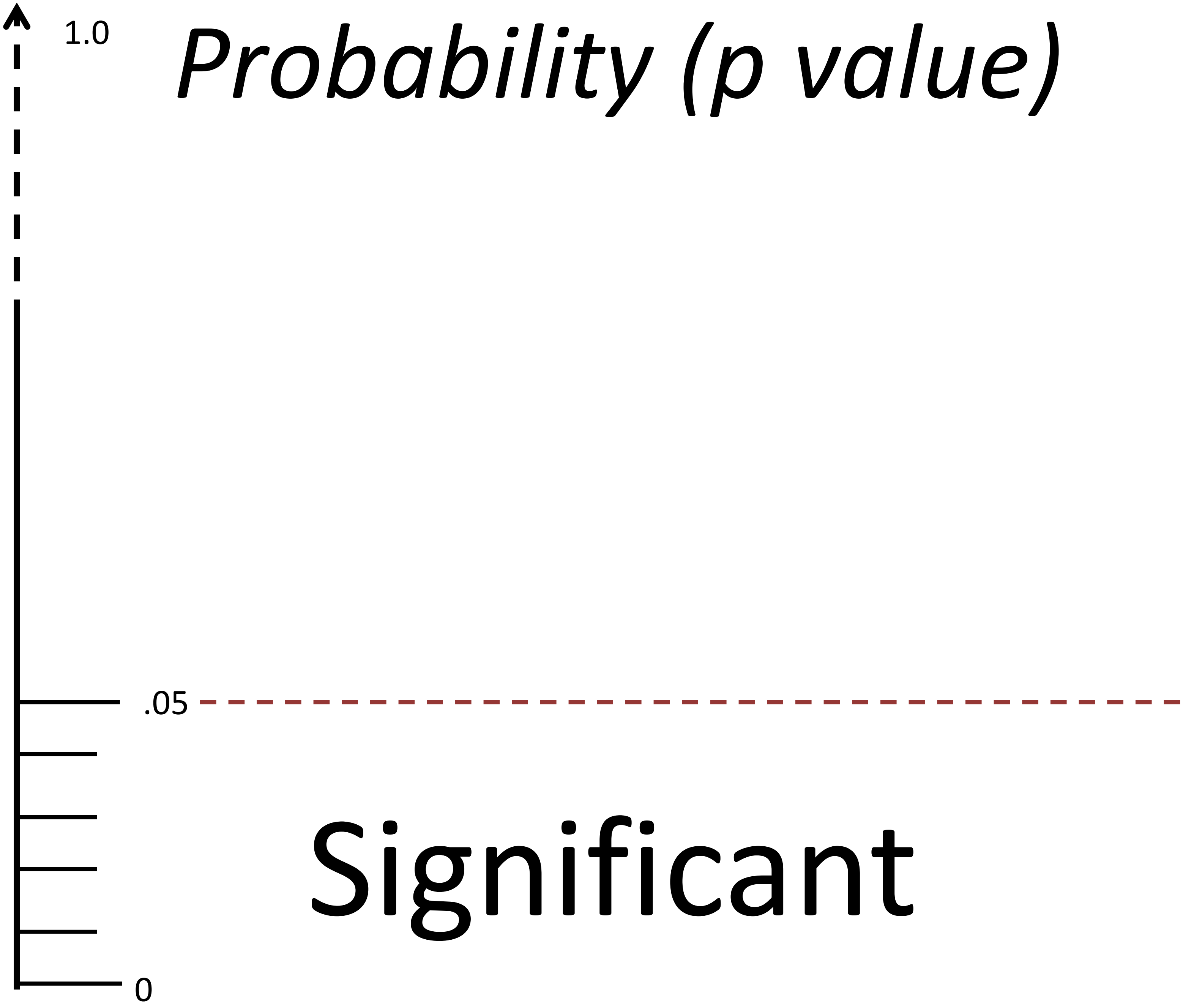
[. . . . . .] Fill the box to void the ticket

How are you most likely to die? The chart below summarizes the probability of death by various causes for the average Canadian.



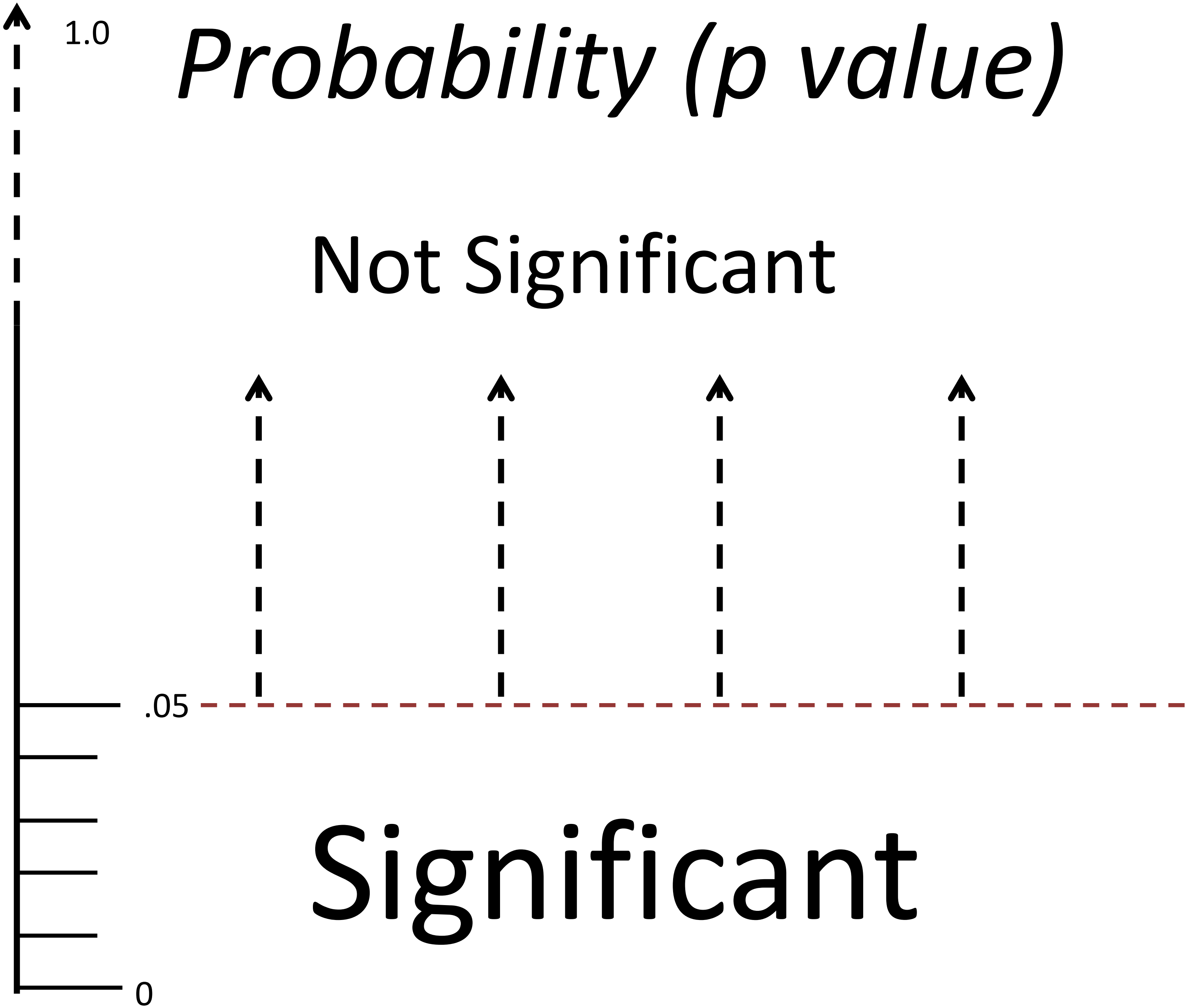
Part 4 Significance Tests & Probability

Probability (p value)



Significant

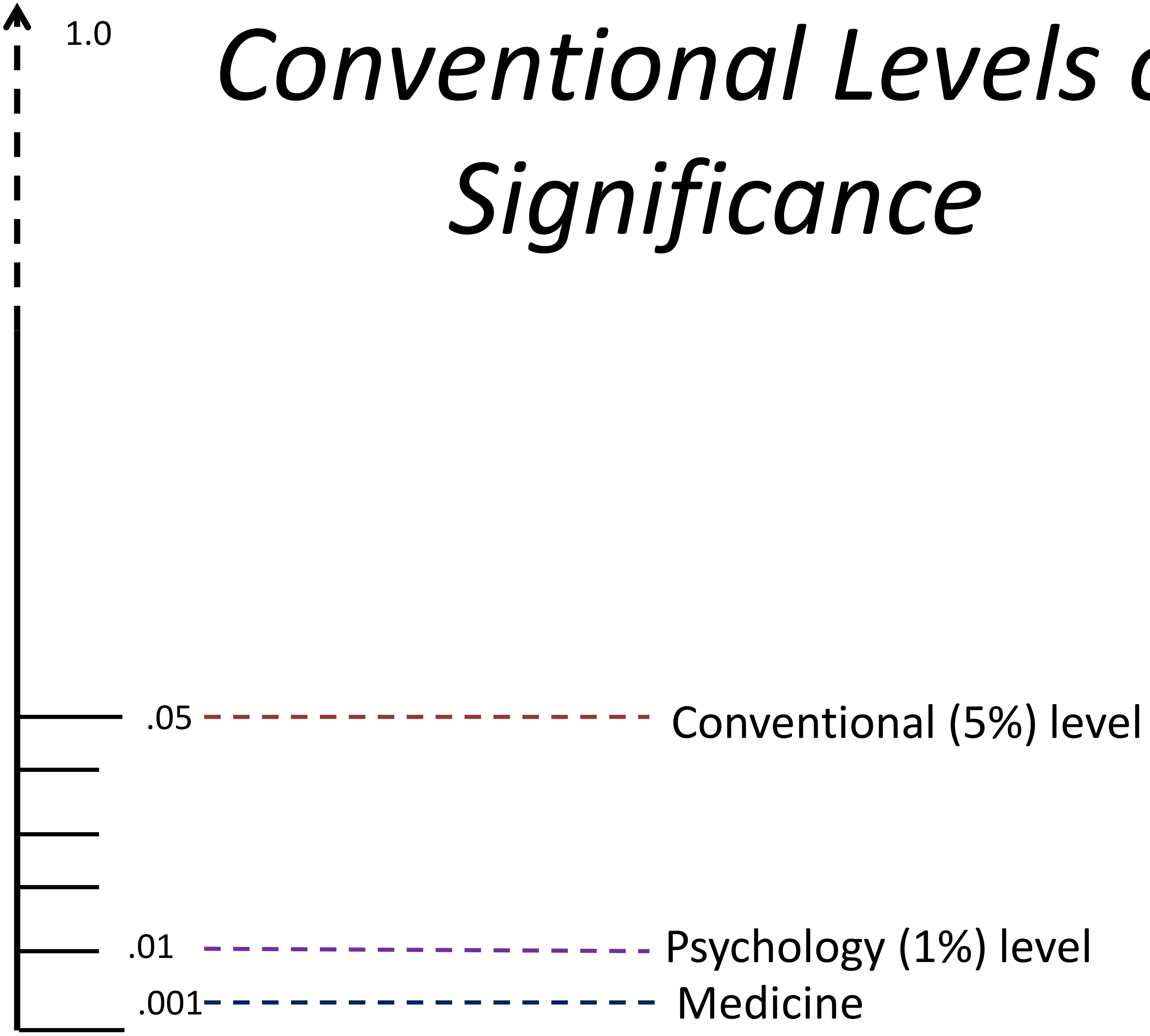
Probability (p value)



Not Significant

Significant

Conventional Levels of Significance



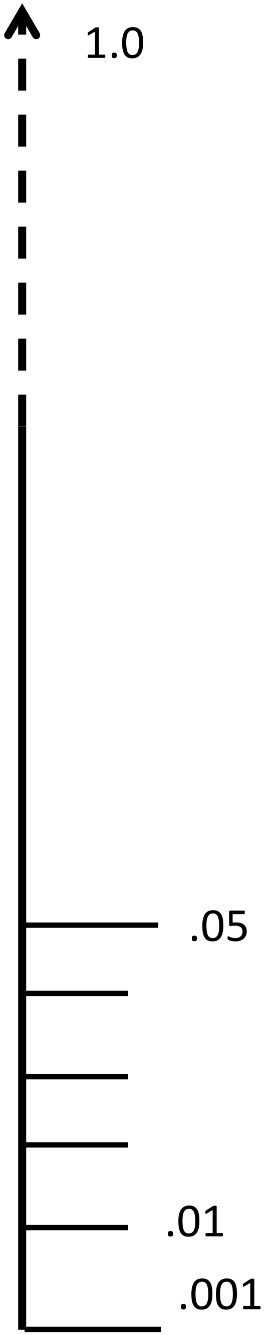
1.0

.05 Conventional (5%) level

.01 Psychology (1%) level

.001 Medicine

Conventional Levels of Significance (Stars)

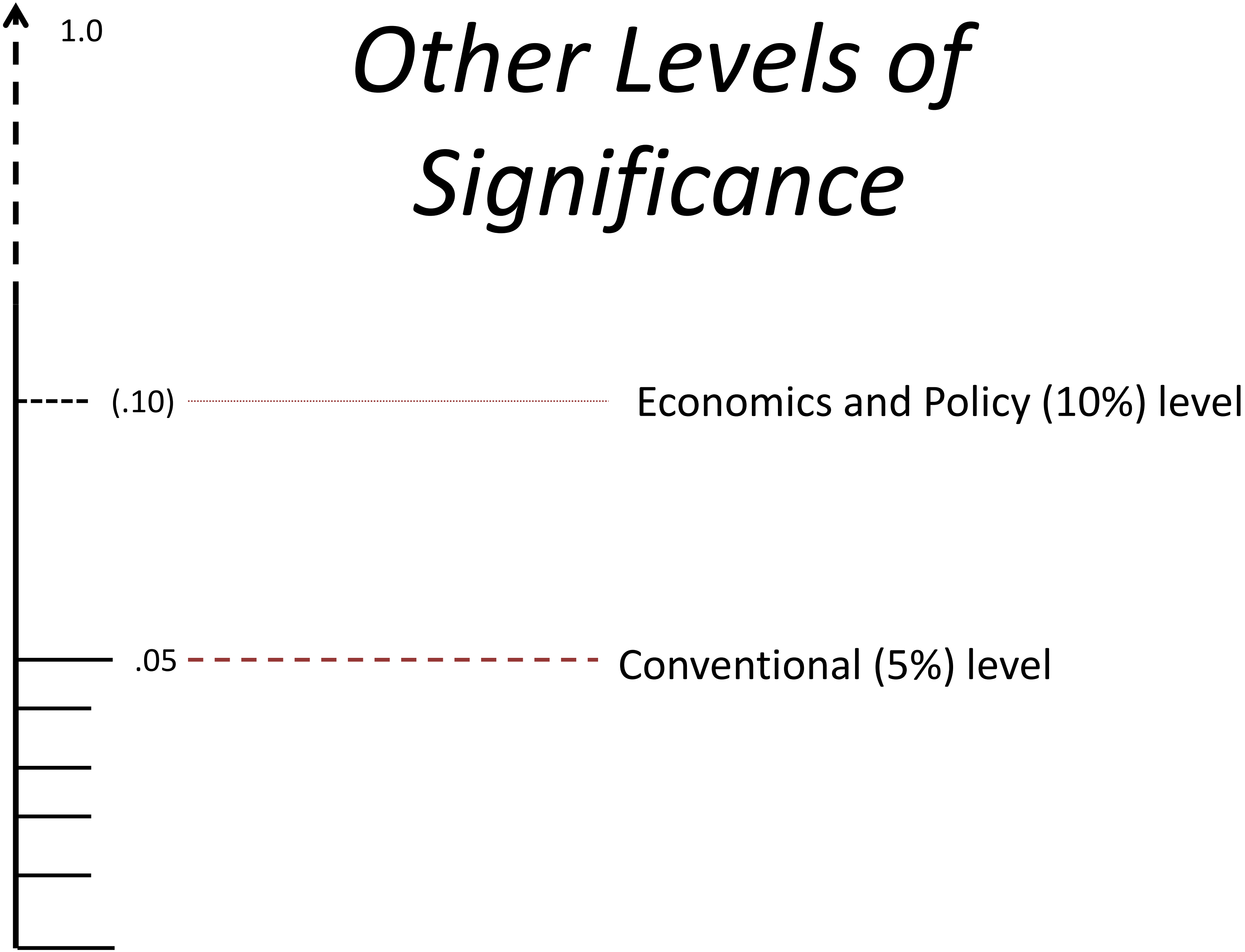


* Conventional (5%) level

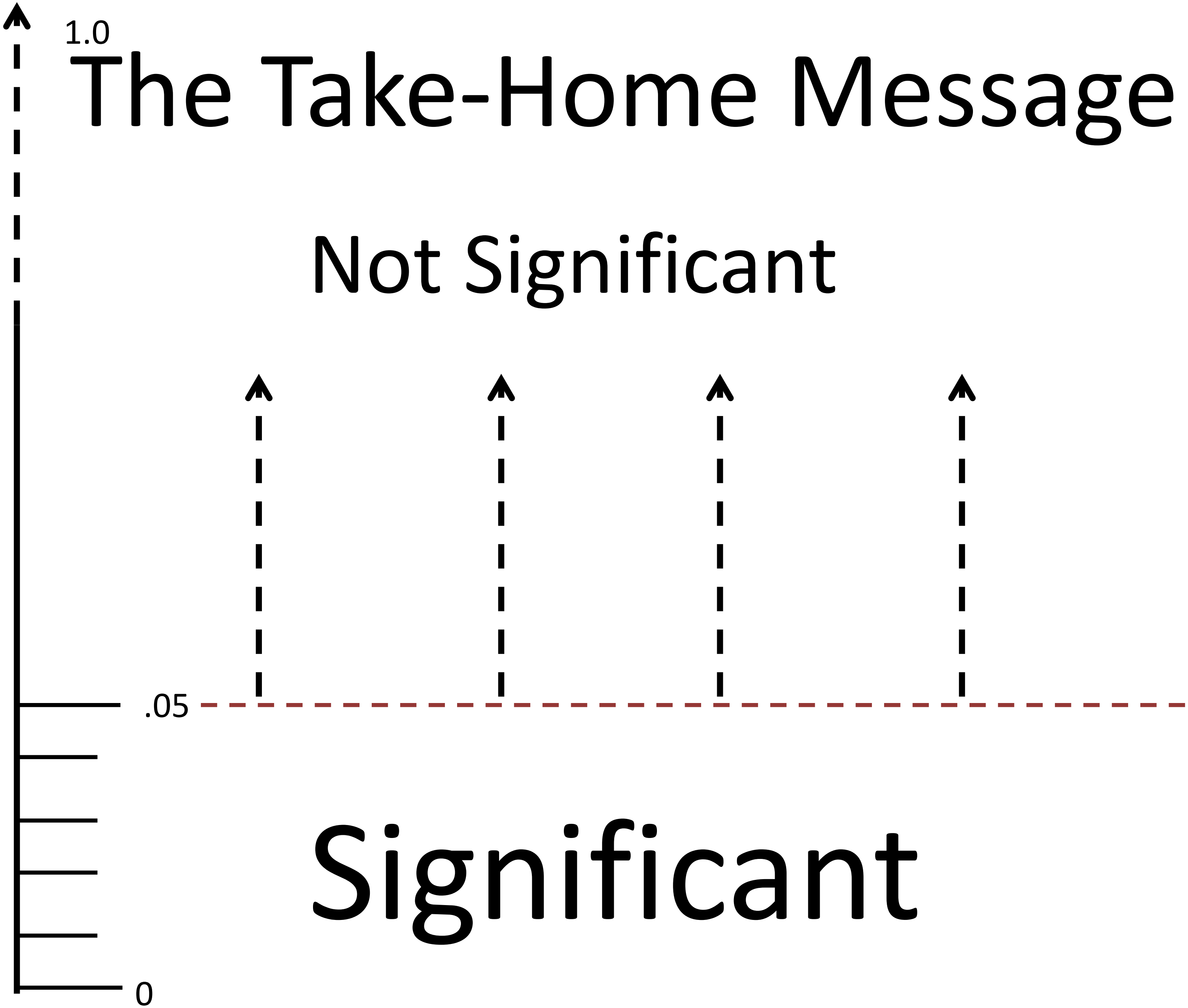
** Psychology (1%) level

*** Medicine

Other Levels of Significance



The Take-Home Message #2

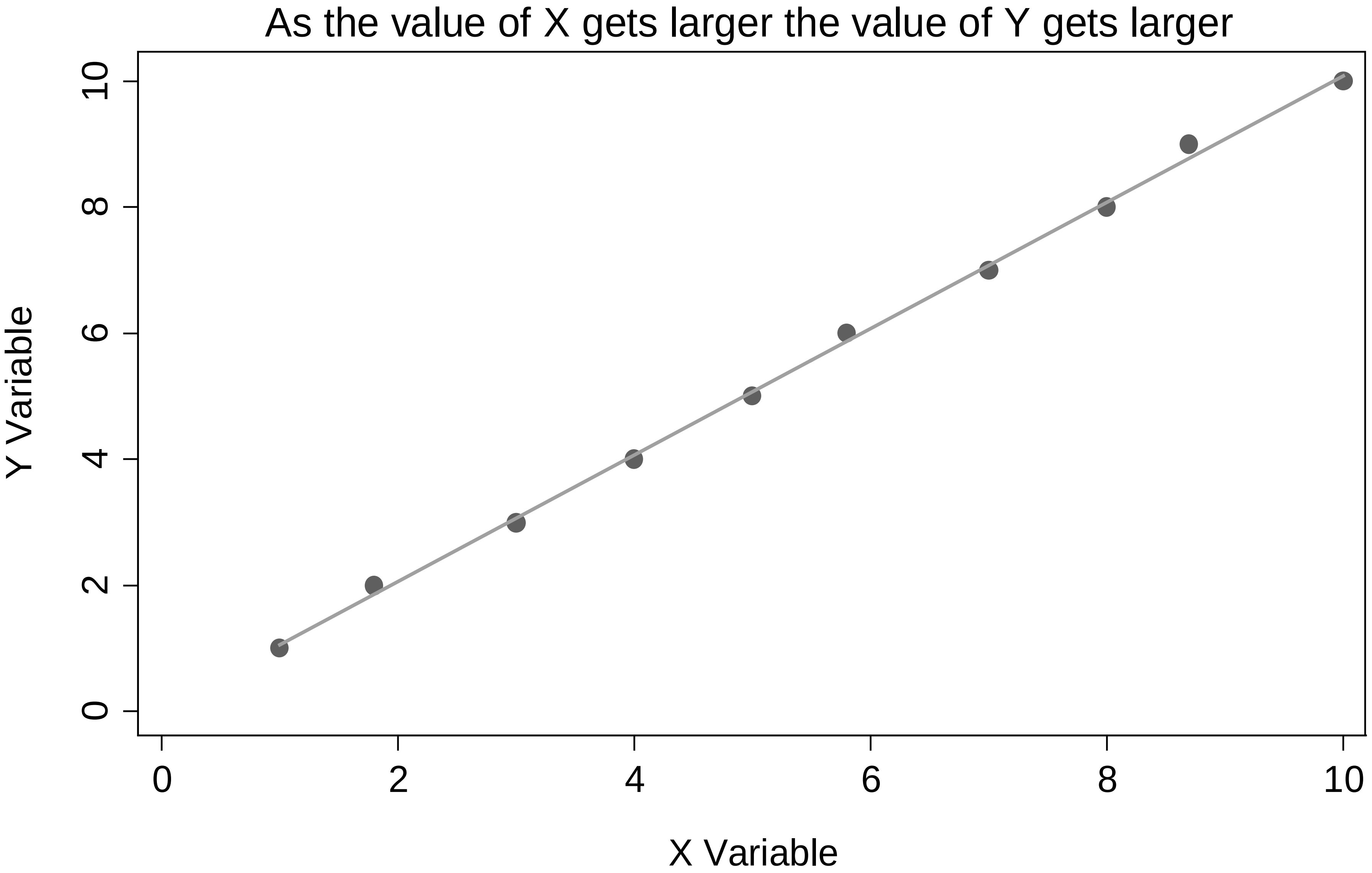


Not Significant

Significant

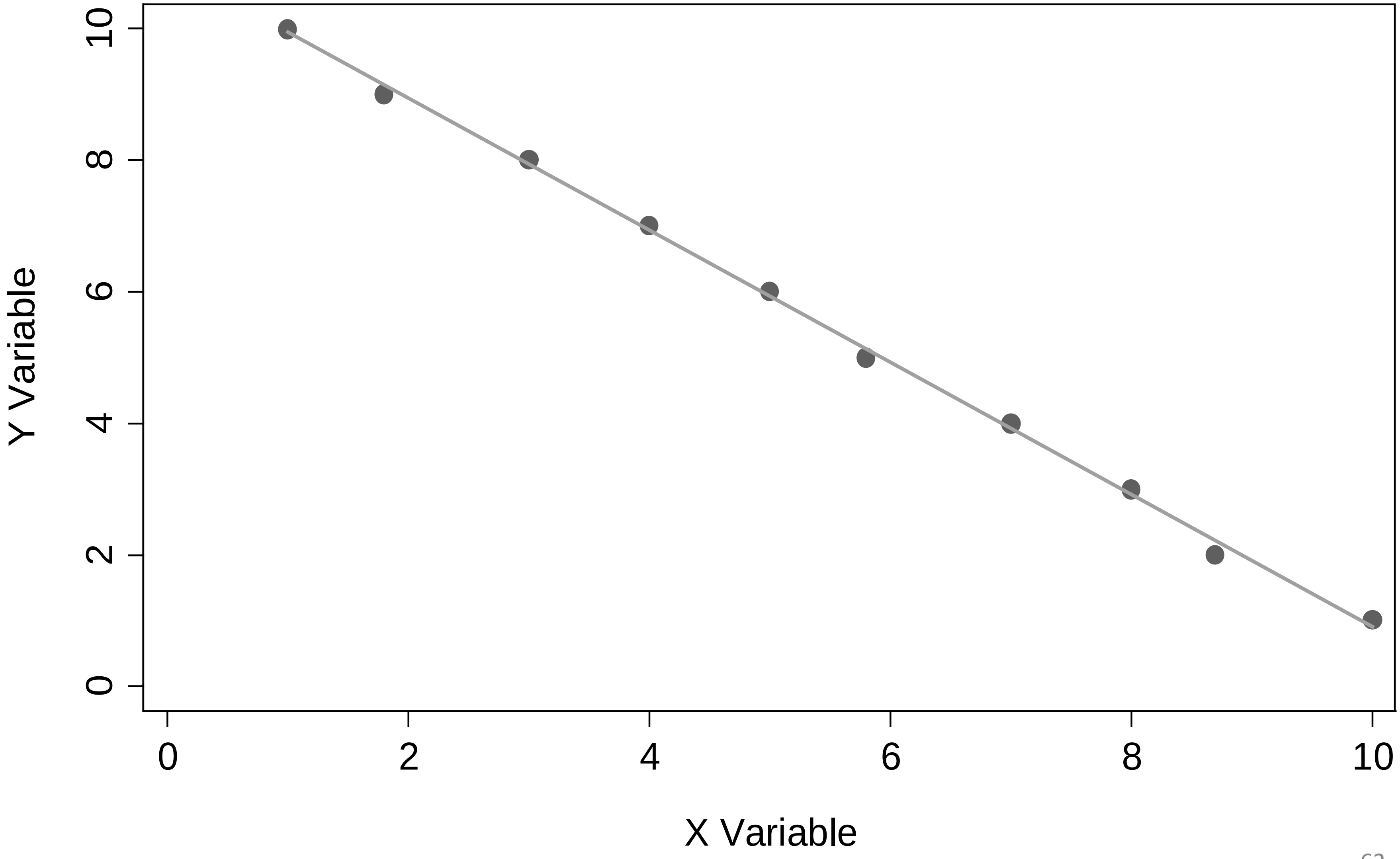
Part 5 Relationships in Data (probably revision)

Positive Relationship



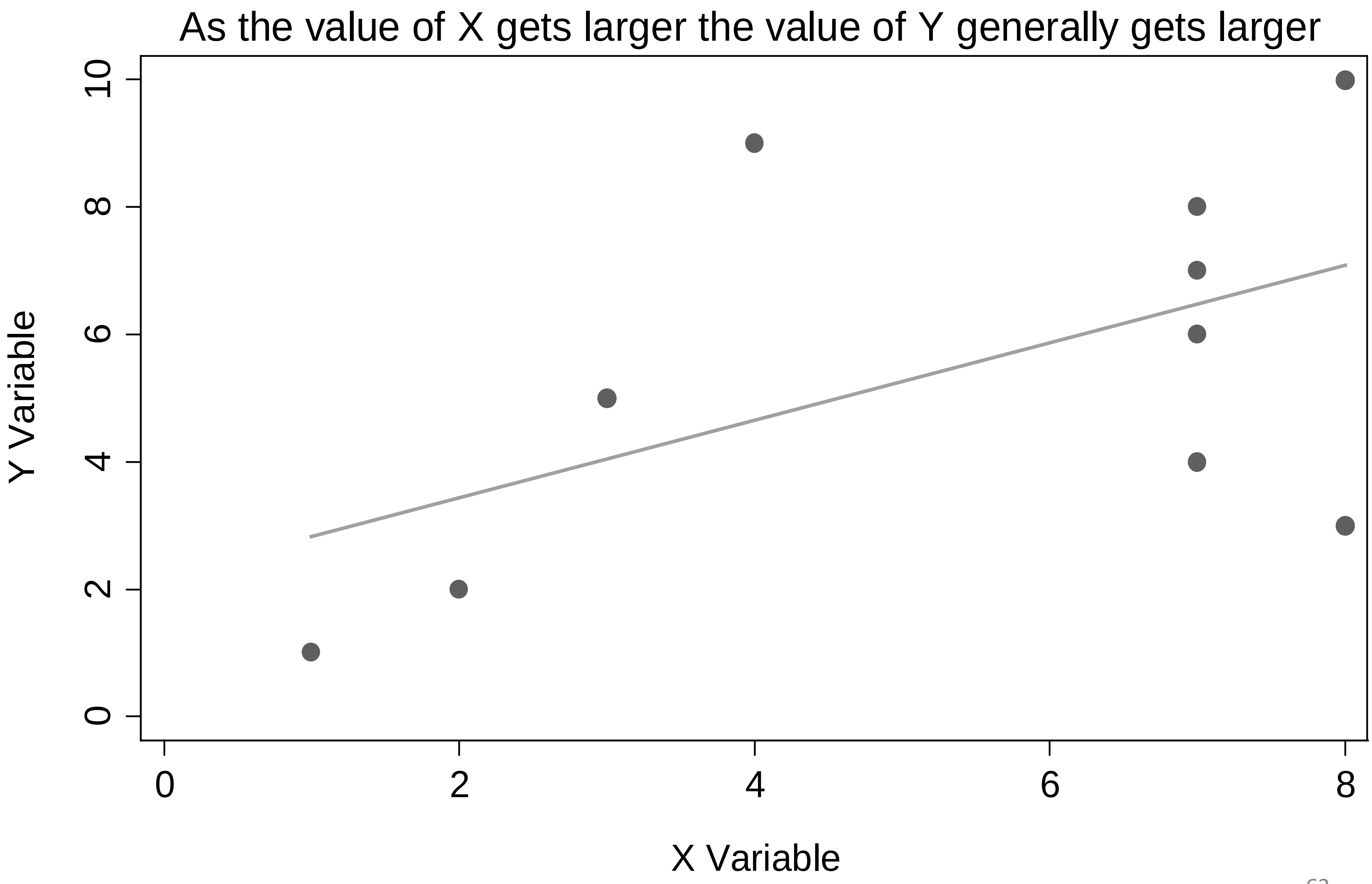
As the value of X gets larger the value of Y gets larger

Negative Relationship

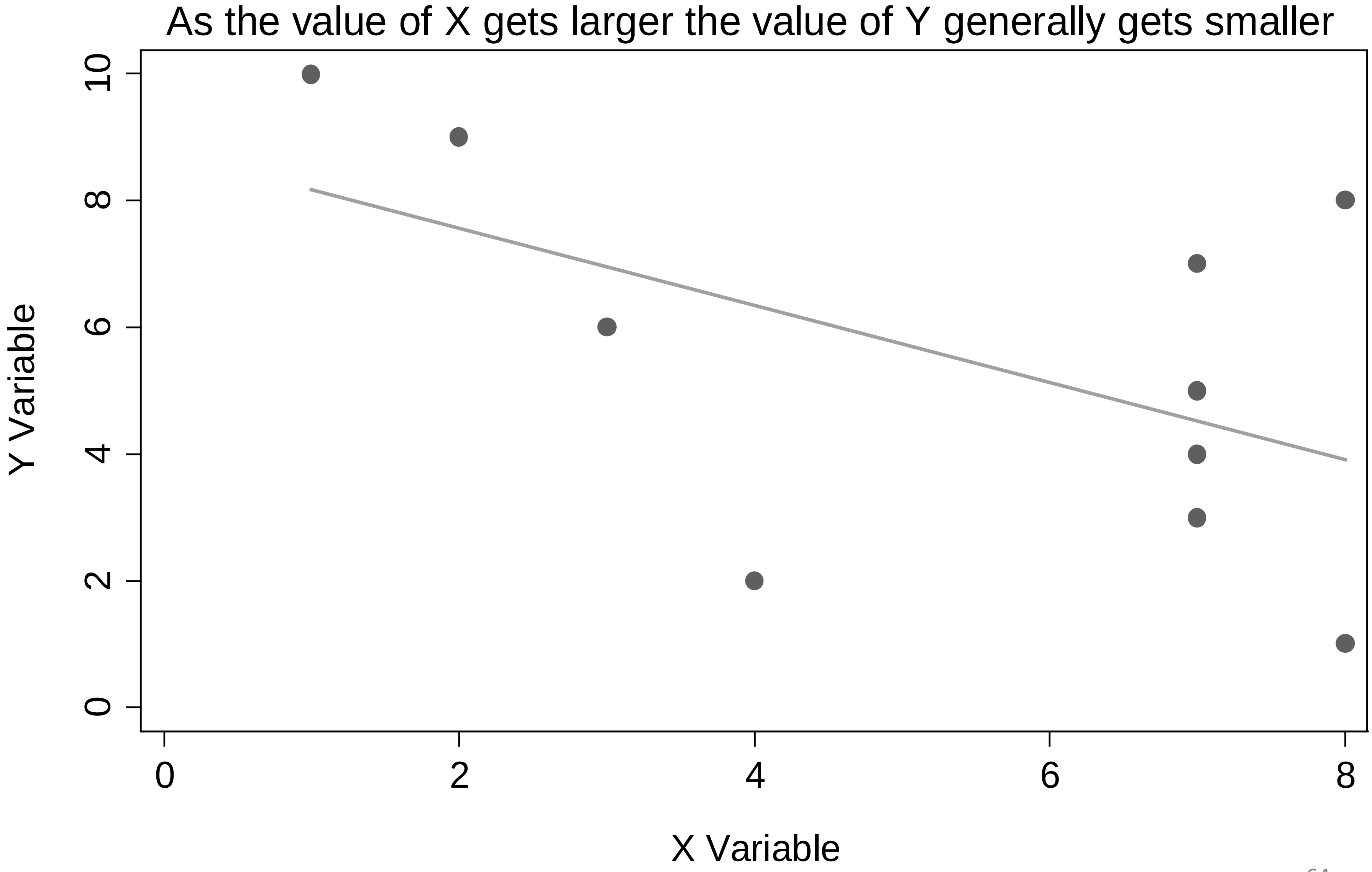


As the value of X gets larger the value of Y gets smaller

Weaker Positive Relationship

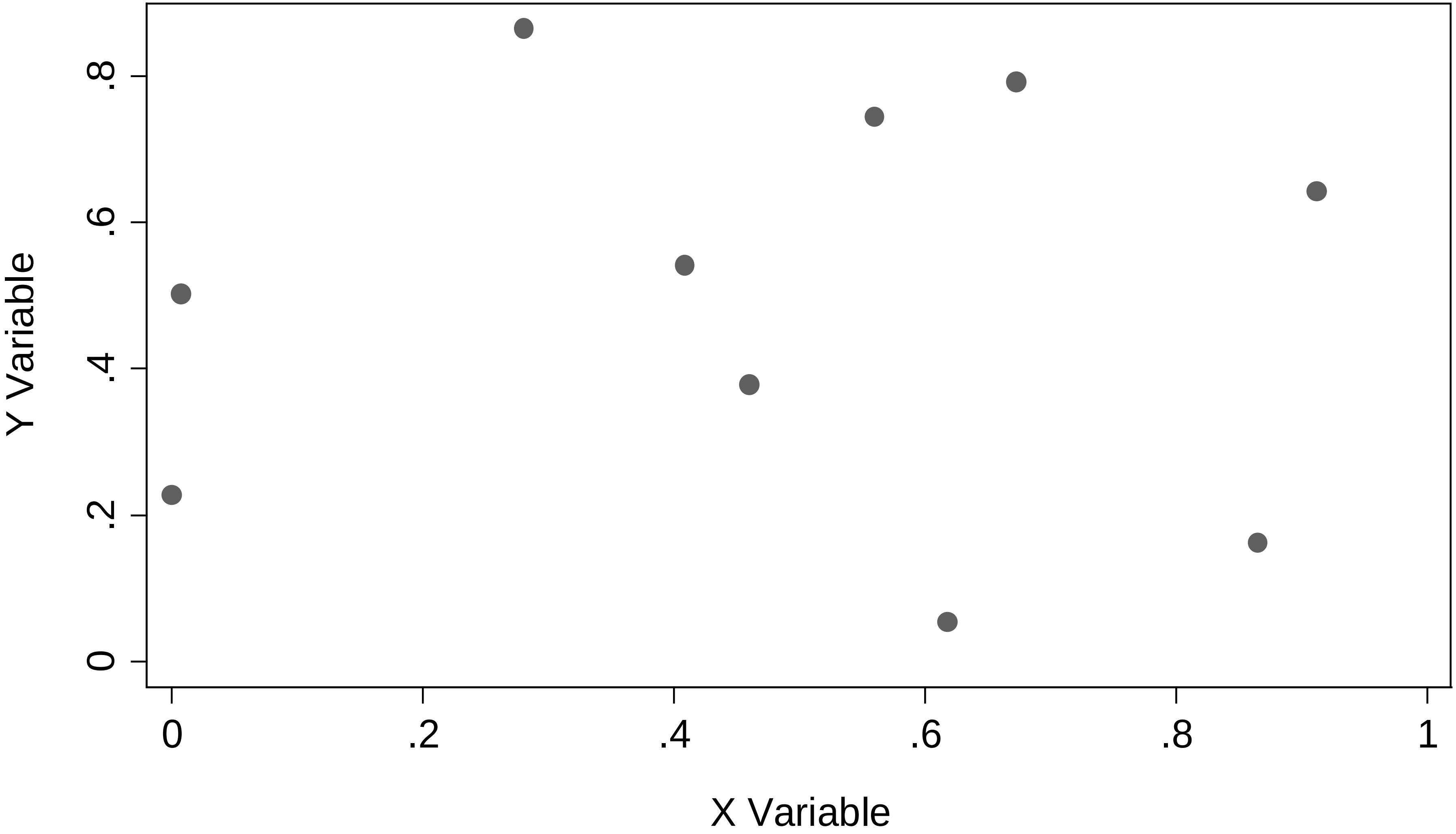


Weaker Negative Relationship



No (linear) Relationship

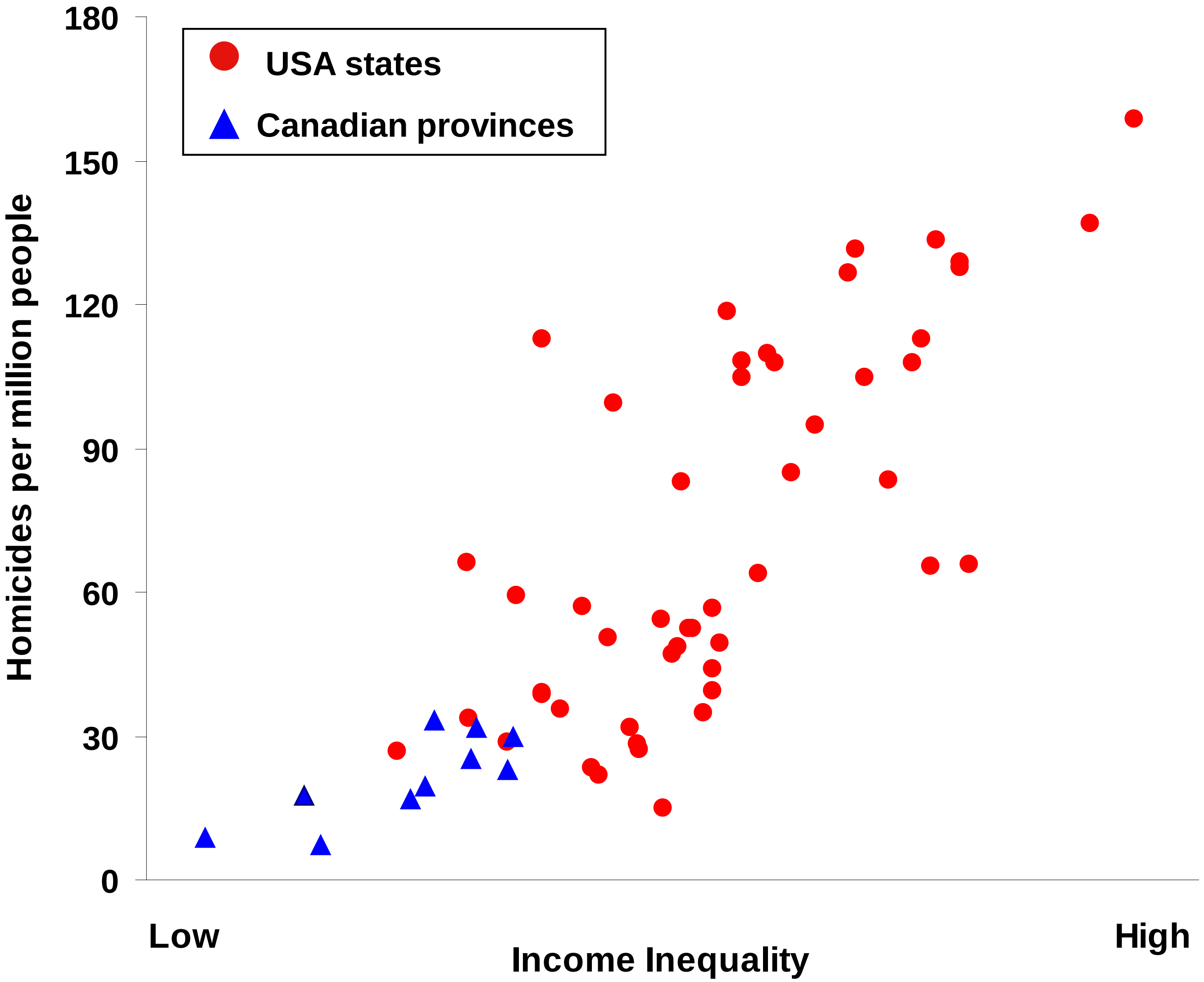
As the value of X gets larger
any guess as to what happens to the value of Y



Activity 3

1. Explain the graph below

Homicide rates are higher in more unequal US states and Canadian provinces



Daly M, Wilson M, Vasdev S. Income inequality and homicide rates in Canada and the United States. *Can J Crim* 2001; 43: 219-36.

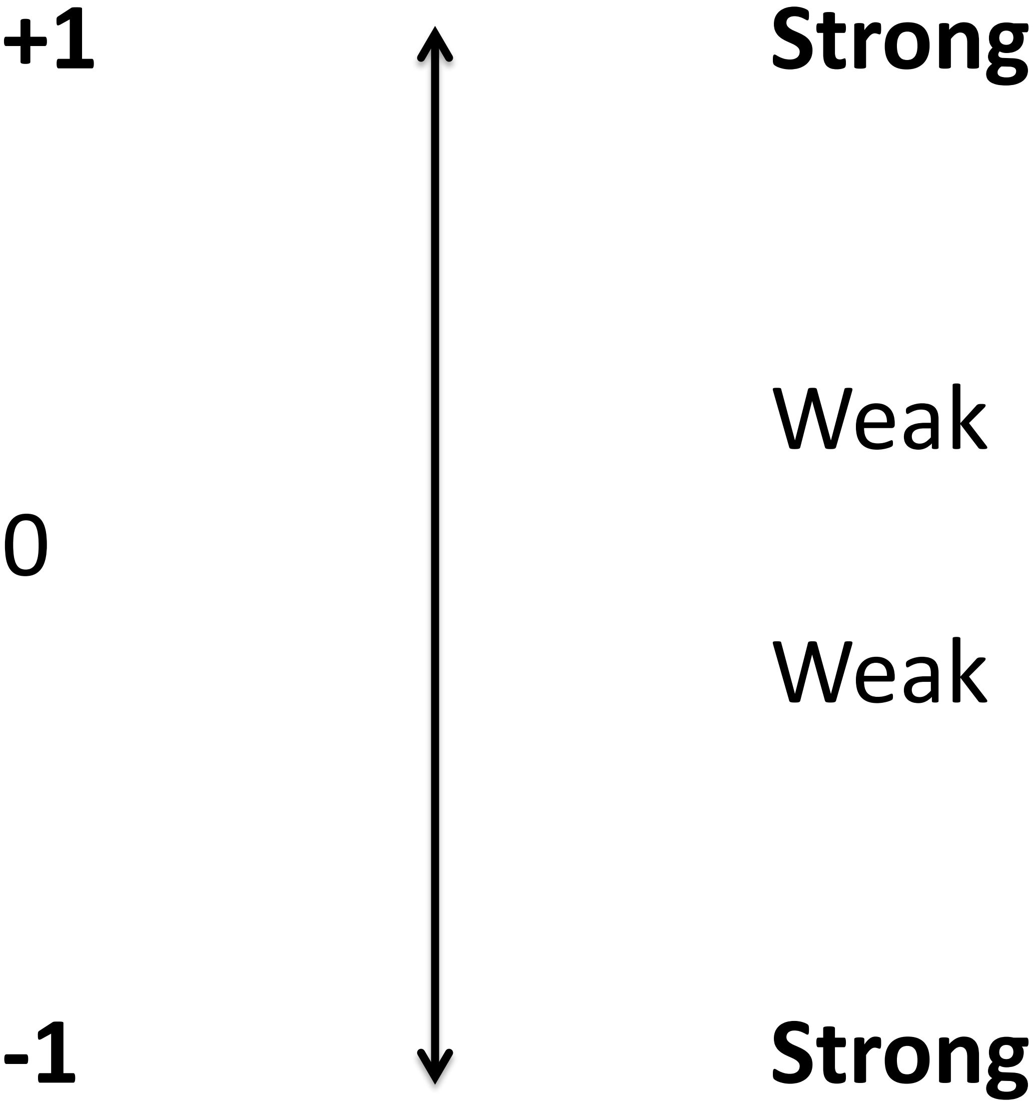
Part 6　Correlations

Vocabulary

Positive (+)

Negative (-)

Pearson’s r



+1 **Strong**

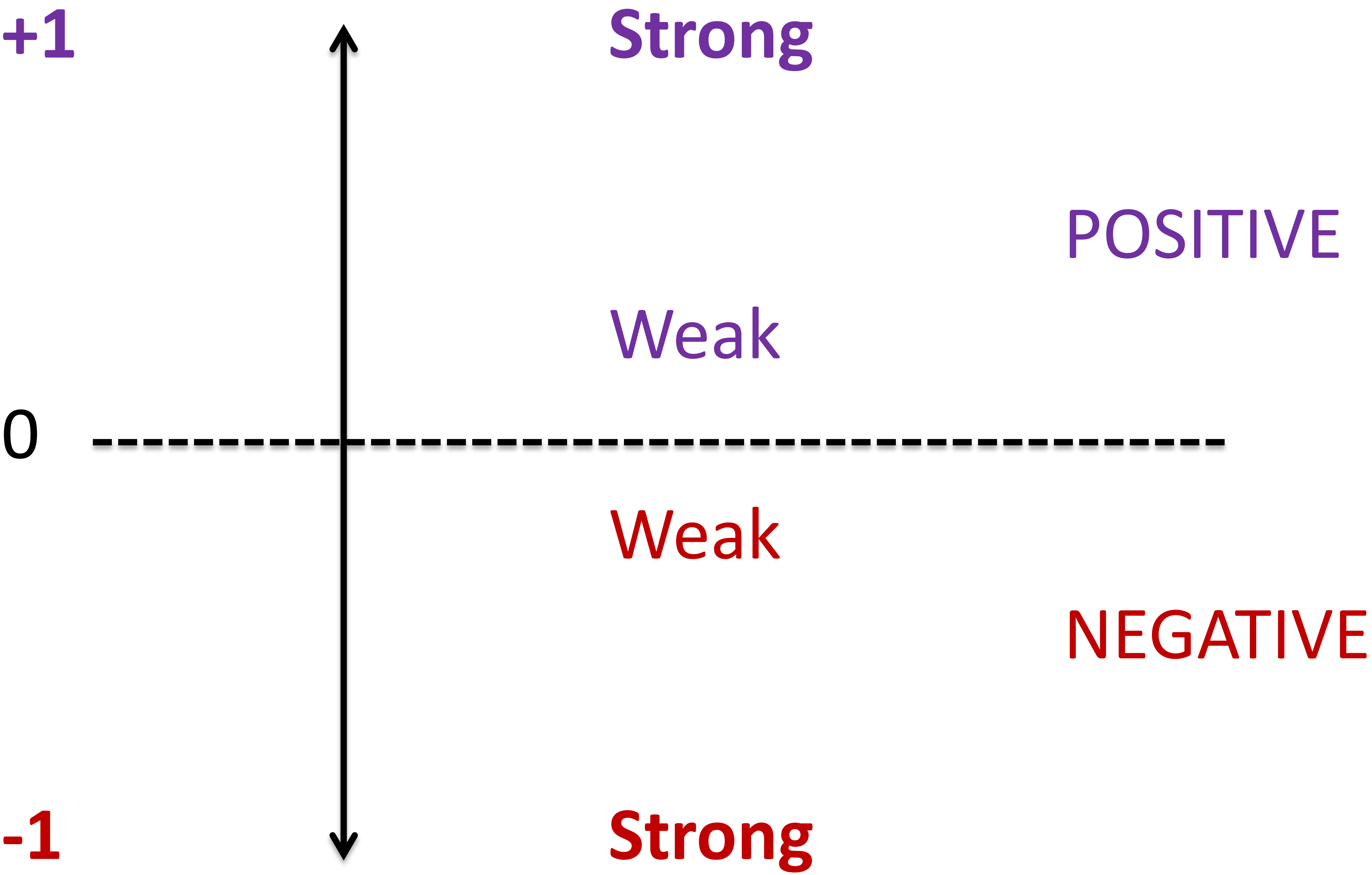
Weak

0

Weak

-1 **Strong**

Pearson’s r



+1

Strong

POSITIVE

Weak

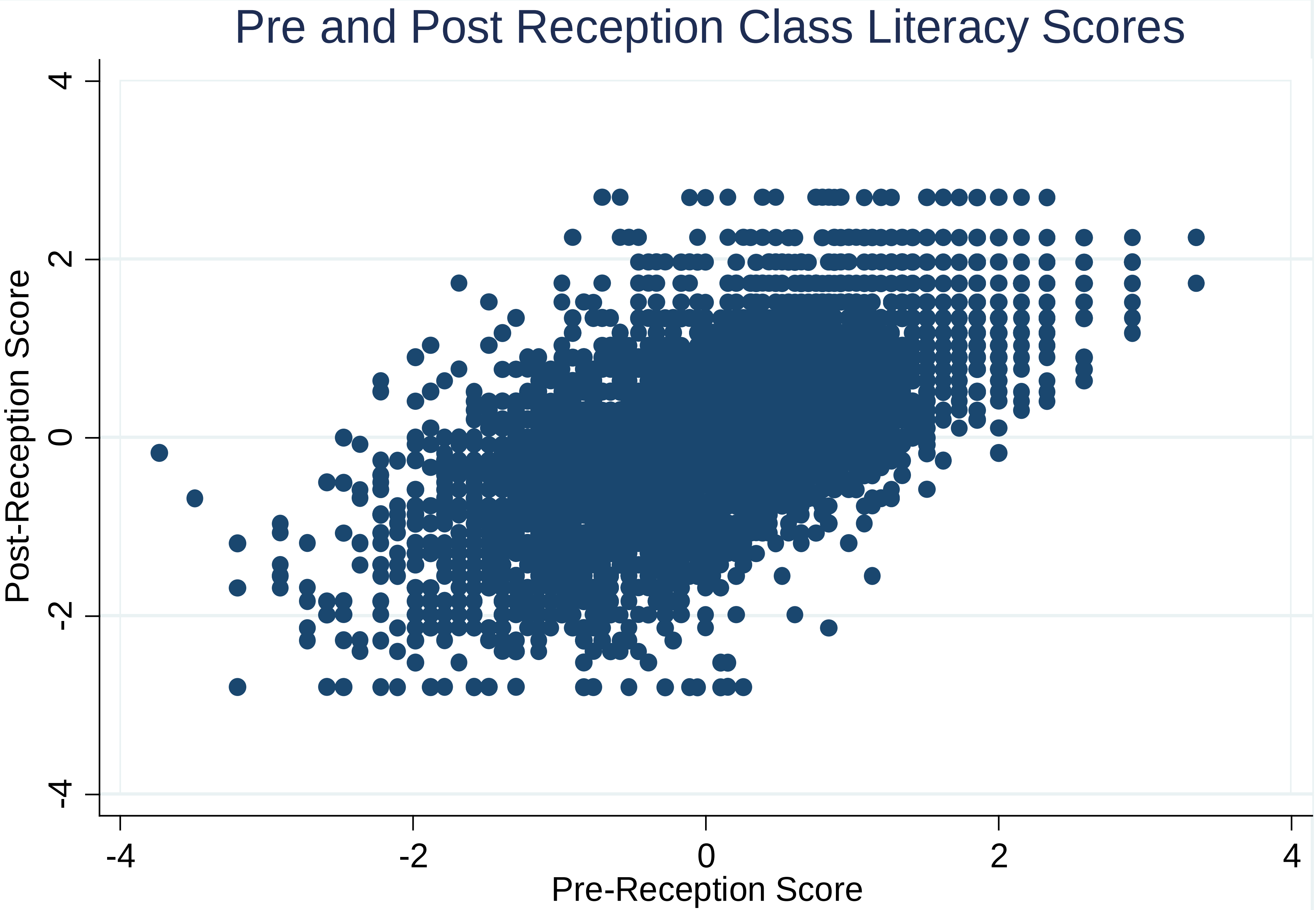
0

Weak

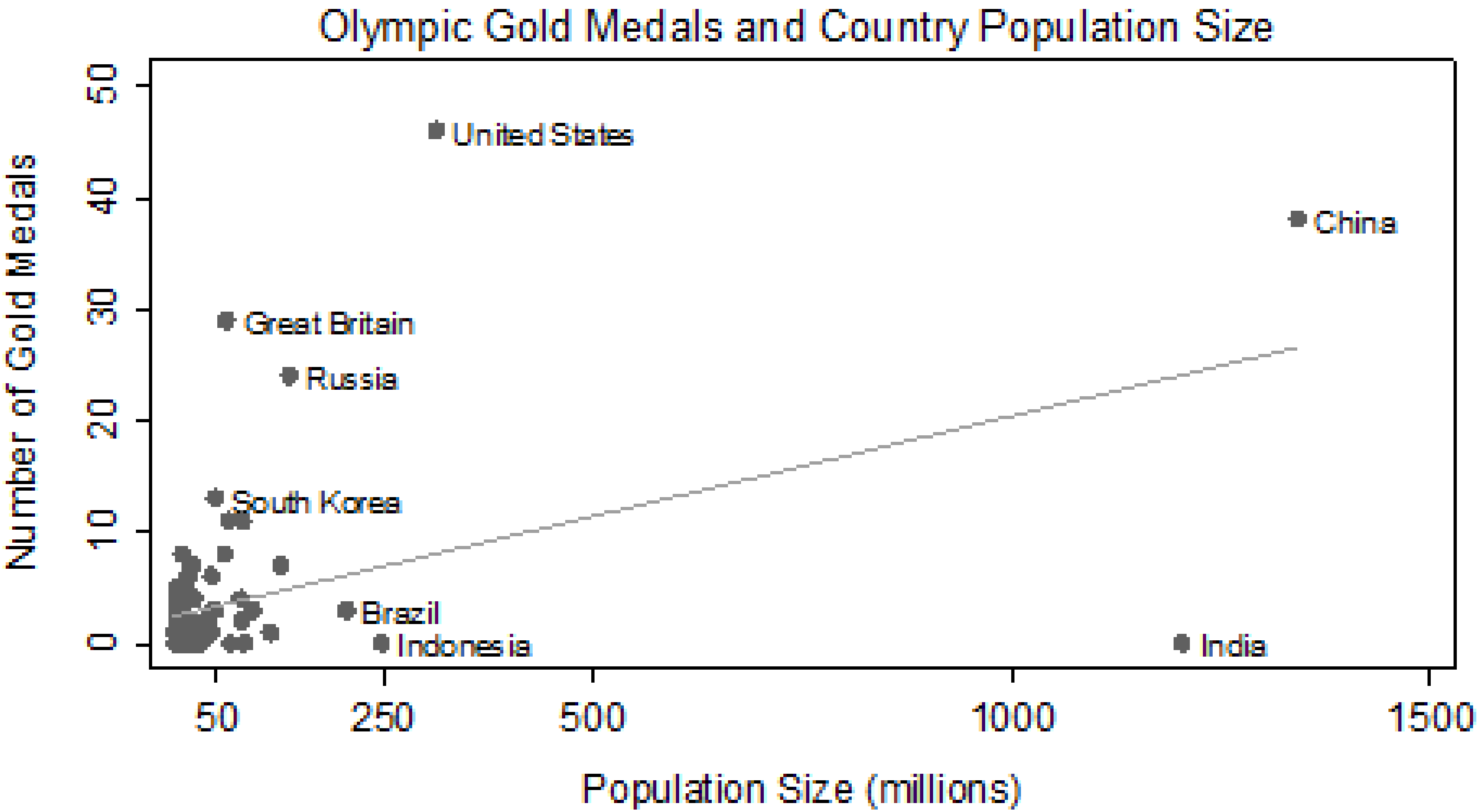
NEGATIVE

-1

Strong



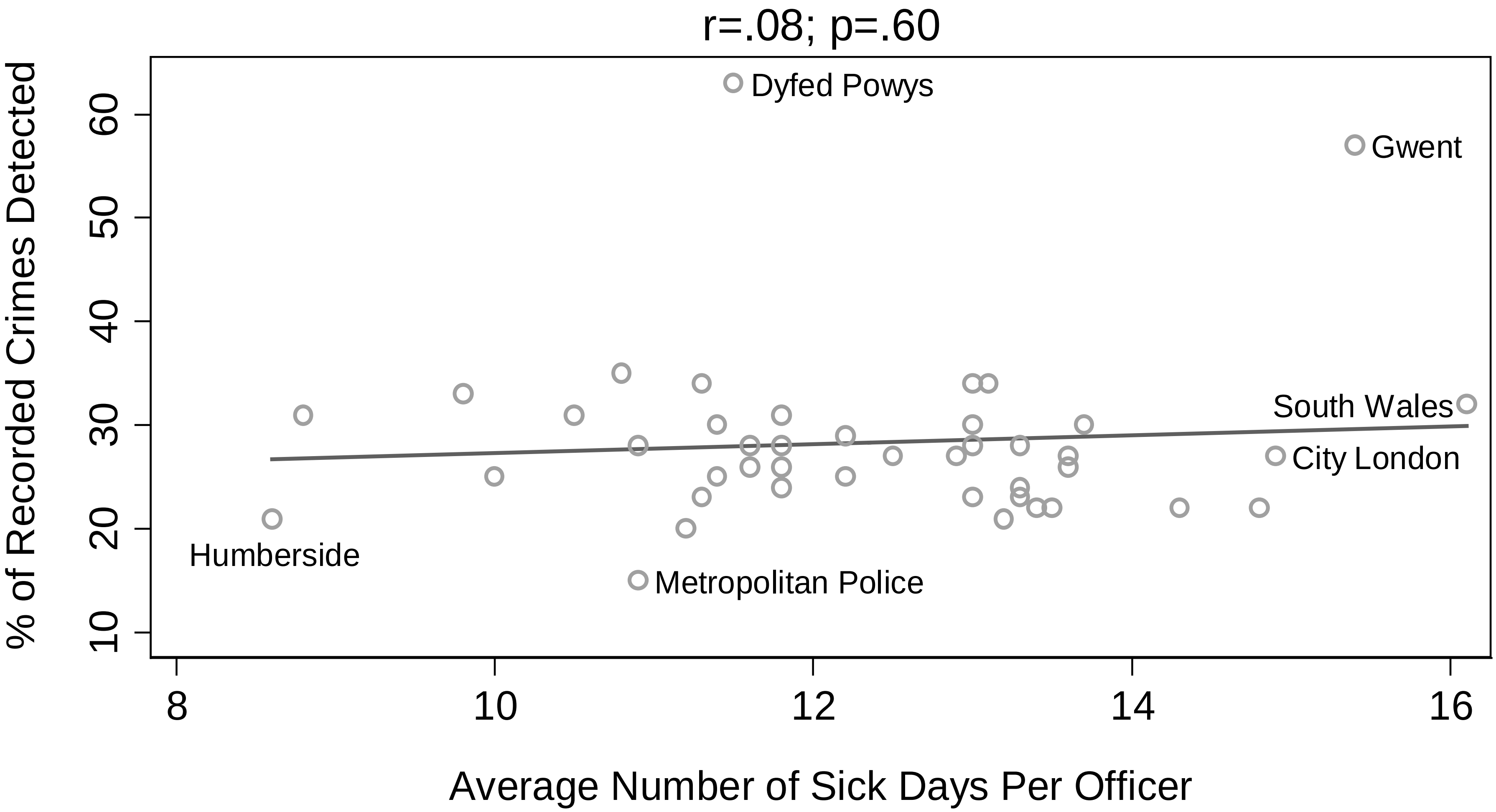
p<.001; r=.66
Source:Blatchford et al (2002), n=4873.



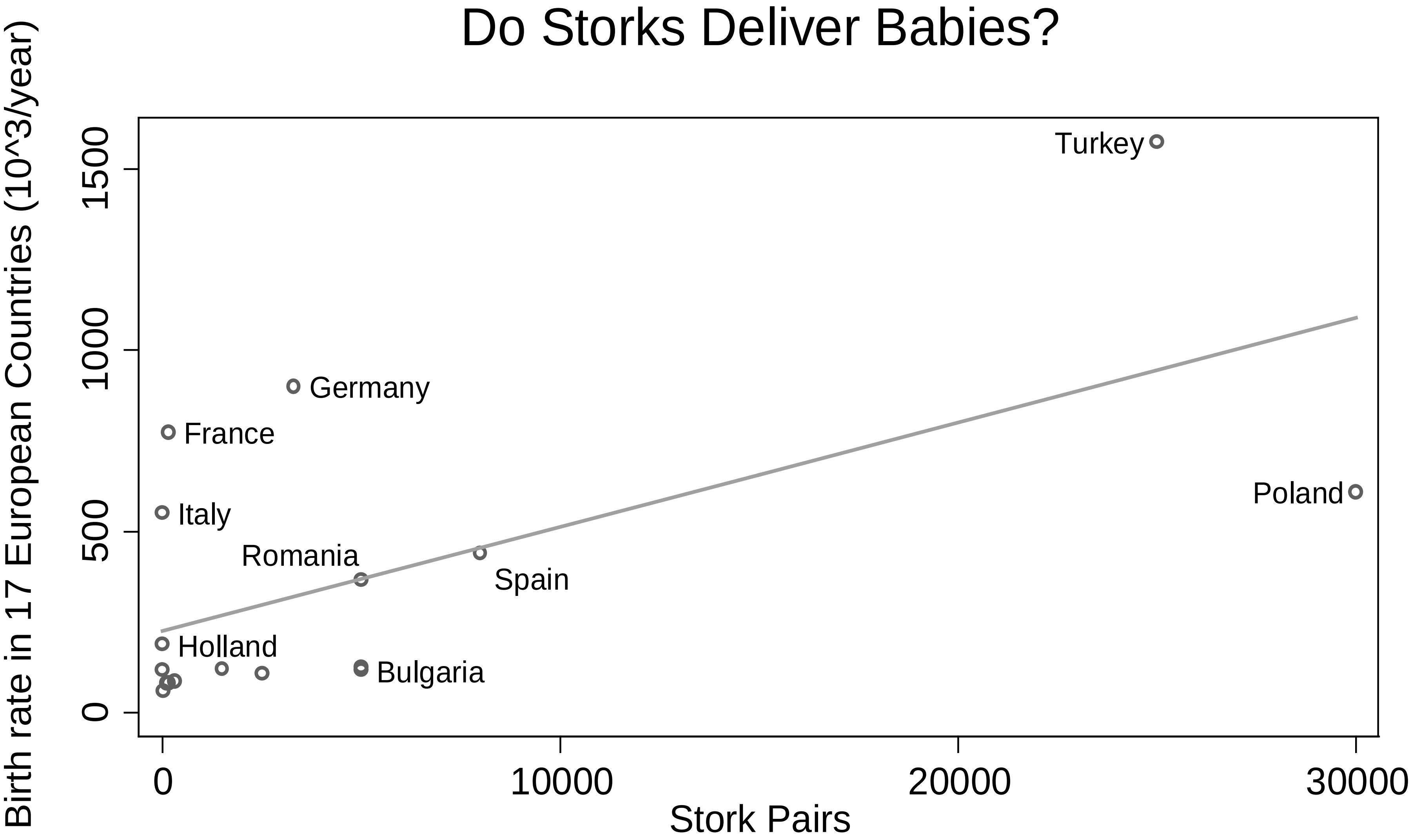
p<.01; r=.46; rsquared=.22;
Data source: Timothy Lethbridge's Blog

Crime Detection and Staff Sickness

Police Forces (England and Wales)



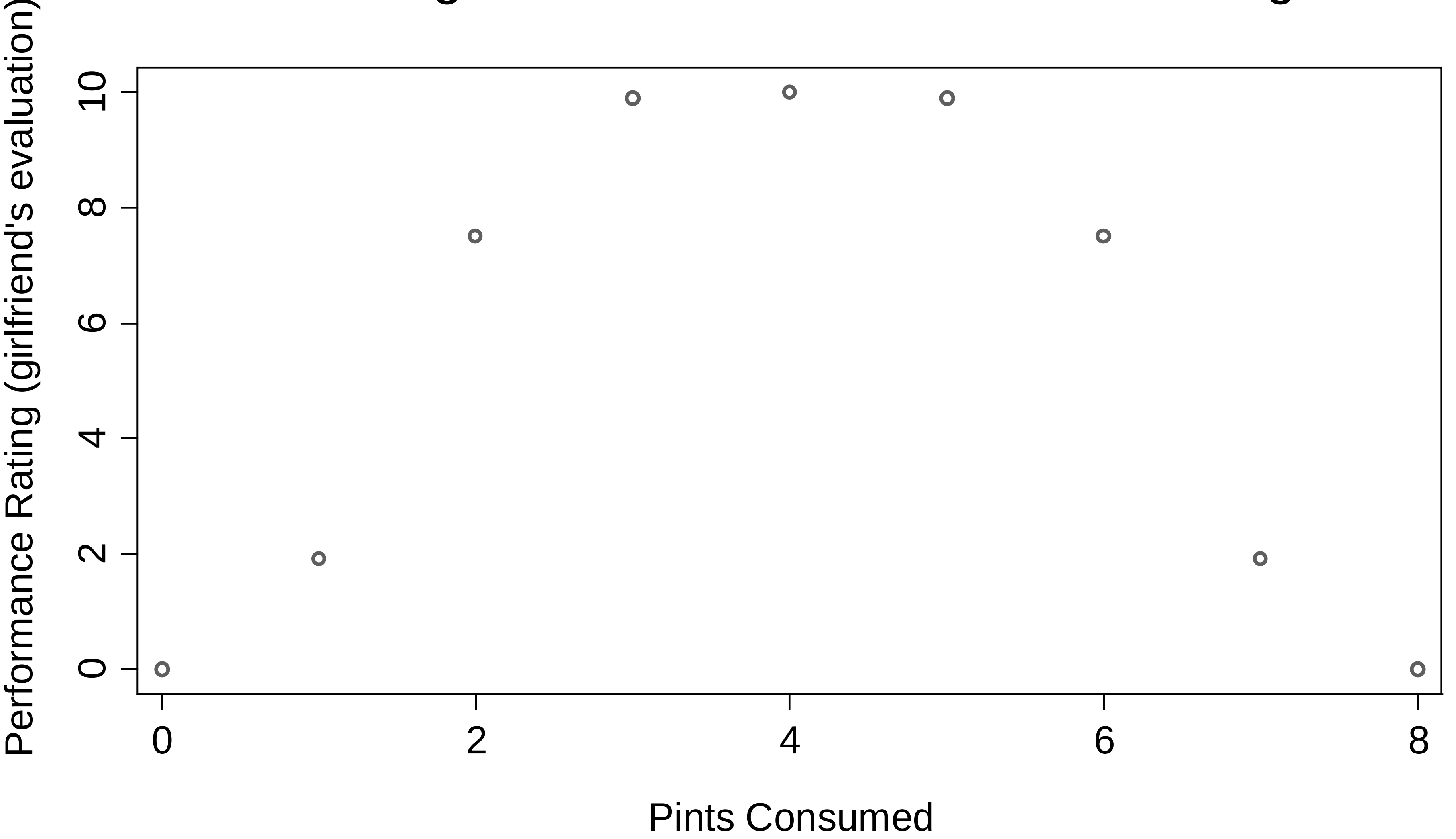
Observer police performance tables
http://observer.theguardian.com/secondterm/table/0,8173,609883,00.html



p=.008; r=.62

Mathews, R. 'Storks Deliver Babies (p=0.008)'
Teaching Statistics. Volume 22, Number 2, Summer 2000

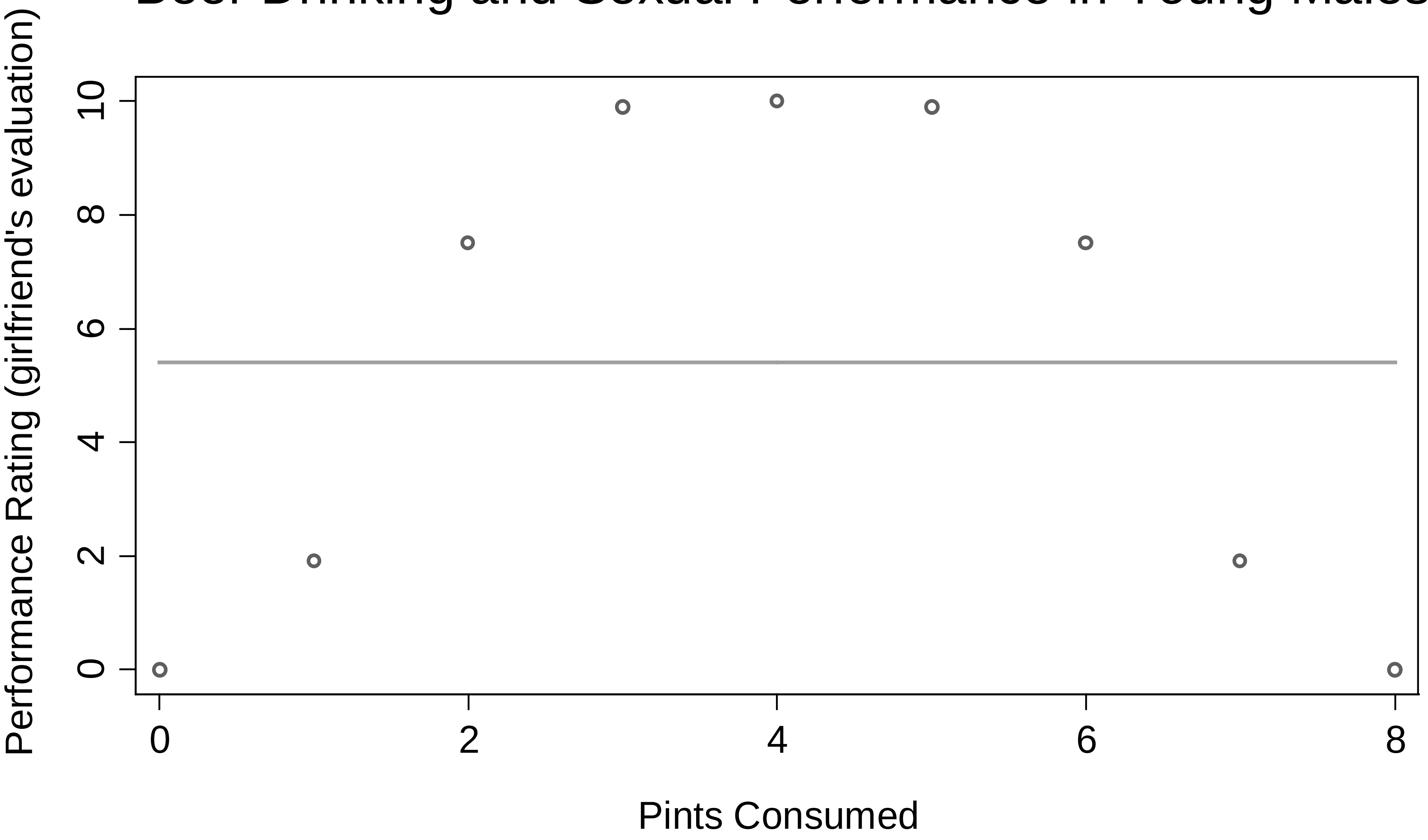
Beer Drinking and Sexual Performance in Young Males



p=1.00; r=.00

Journal of Unethical Studies

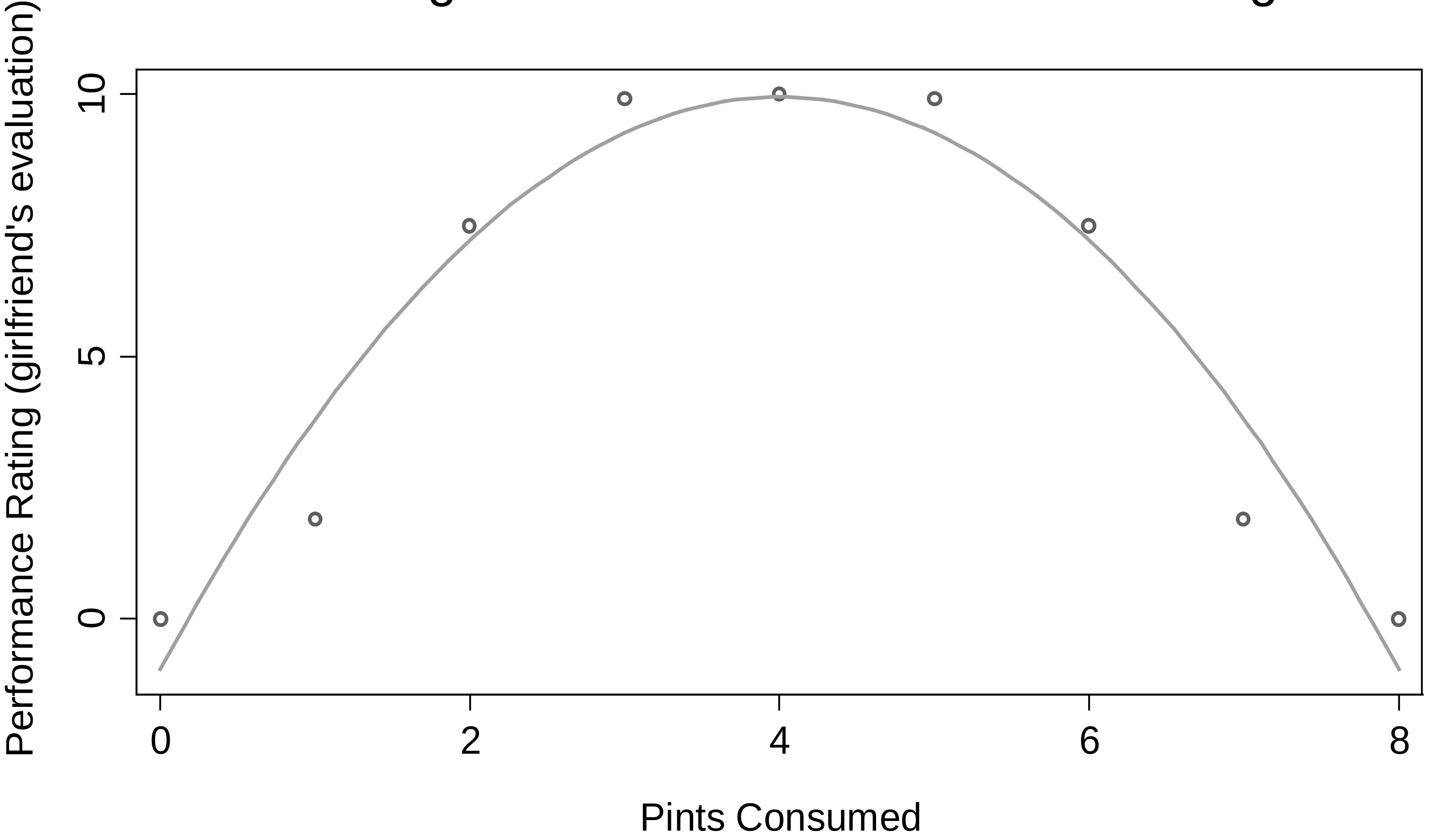
Beer Drinking and Sexual Performance in Young Males



p=1.00; r=.00

Journal of Unethical Studies

Beer Drinking and Sexual Performance in Young Males



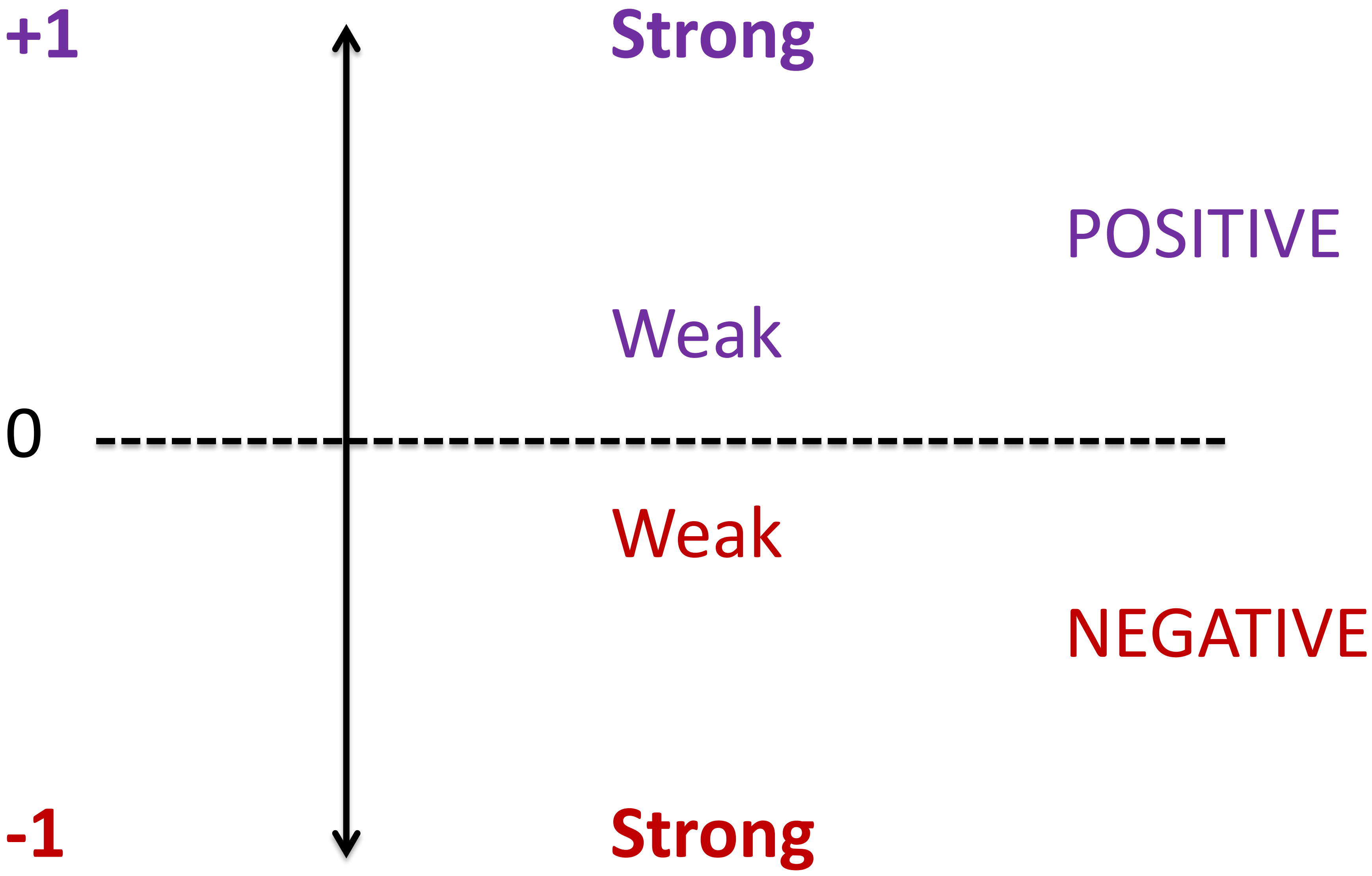
p=1.00; r=.00

Journal of Unethical Studies

Terminology

- Positive or Negative
- Weak of Strong
- Linear relationship
- No linear relationship
- Bivariate regression line (line of best fit)
- Spurious correlation
- Non-linear relationship

Spearman's Rho ρ



+1

Strong

POSITIVE

Weak

0

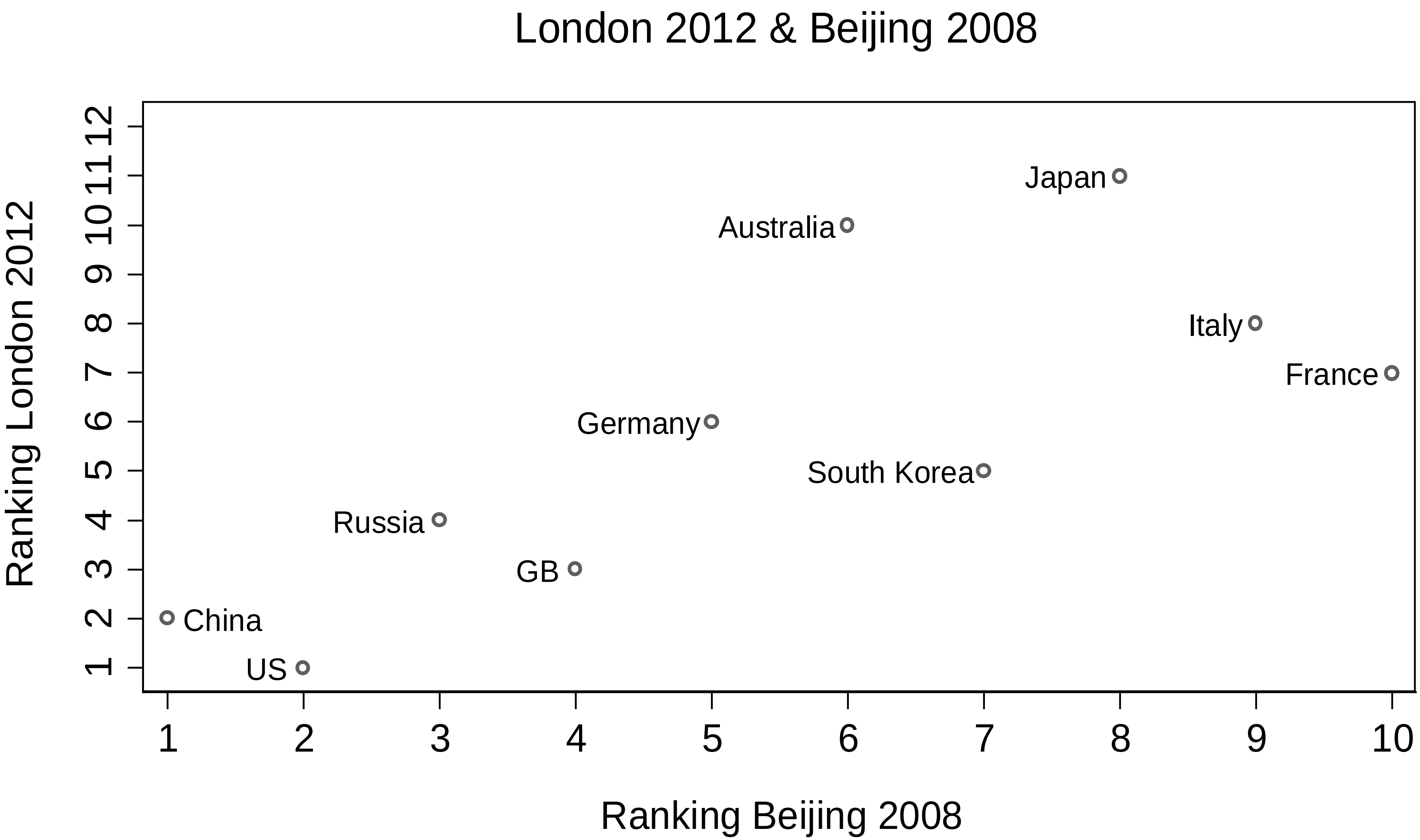
Weak

NEGATIVE

-1

Strong

Olympic Medals Table Rank



p<.01; rho=.81. Source: BBC Sport

Coefficient of Determination

- r^2 takes of values between 0 and 1
- Proportion of variation in Y explained by X

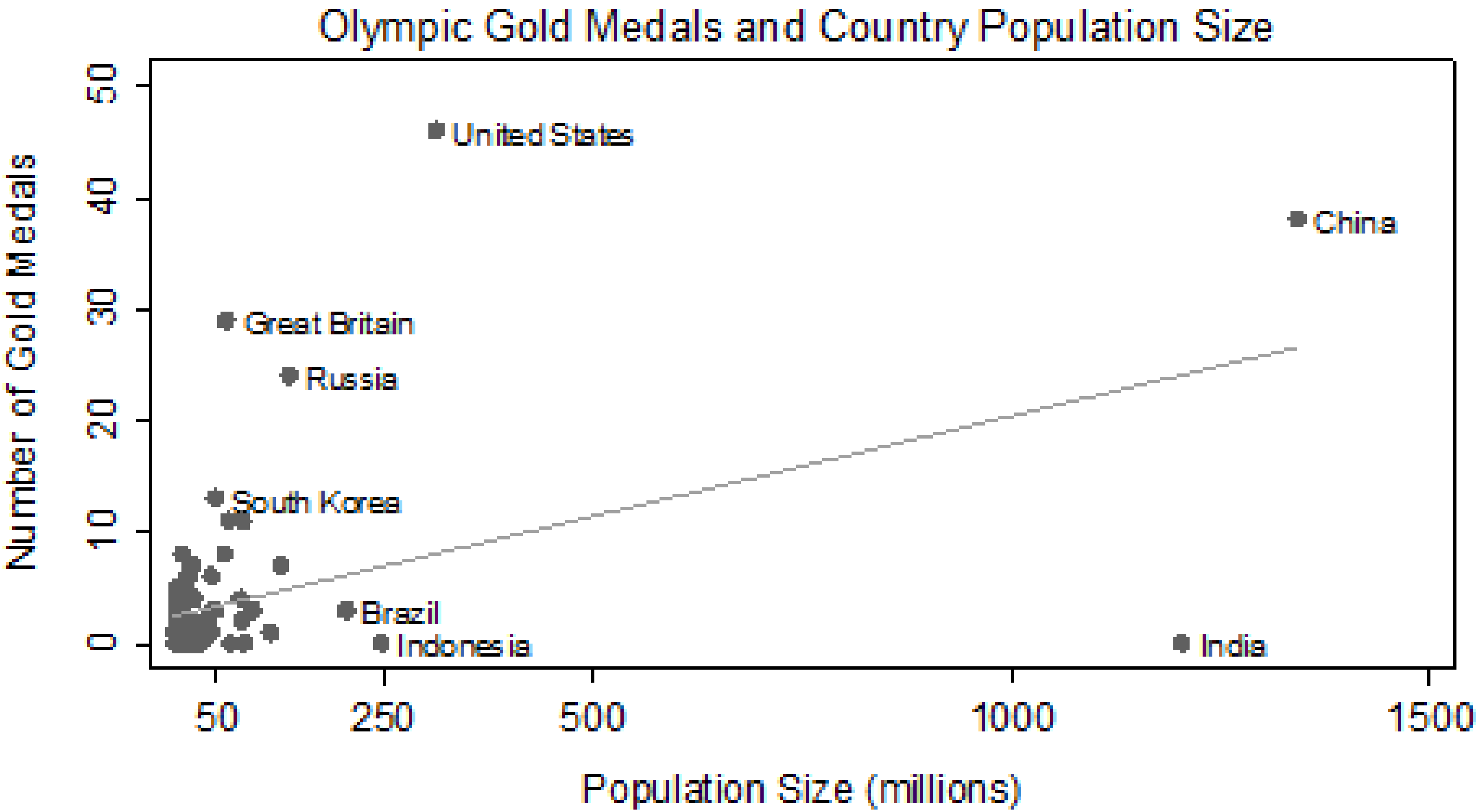
Therefore…

A strong positive correlation $r = .9$ then $r^2 = .81$

A strong negative correlation $r = -.9$ then $r^2 = .81$

Activity 4

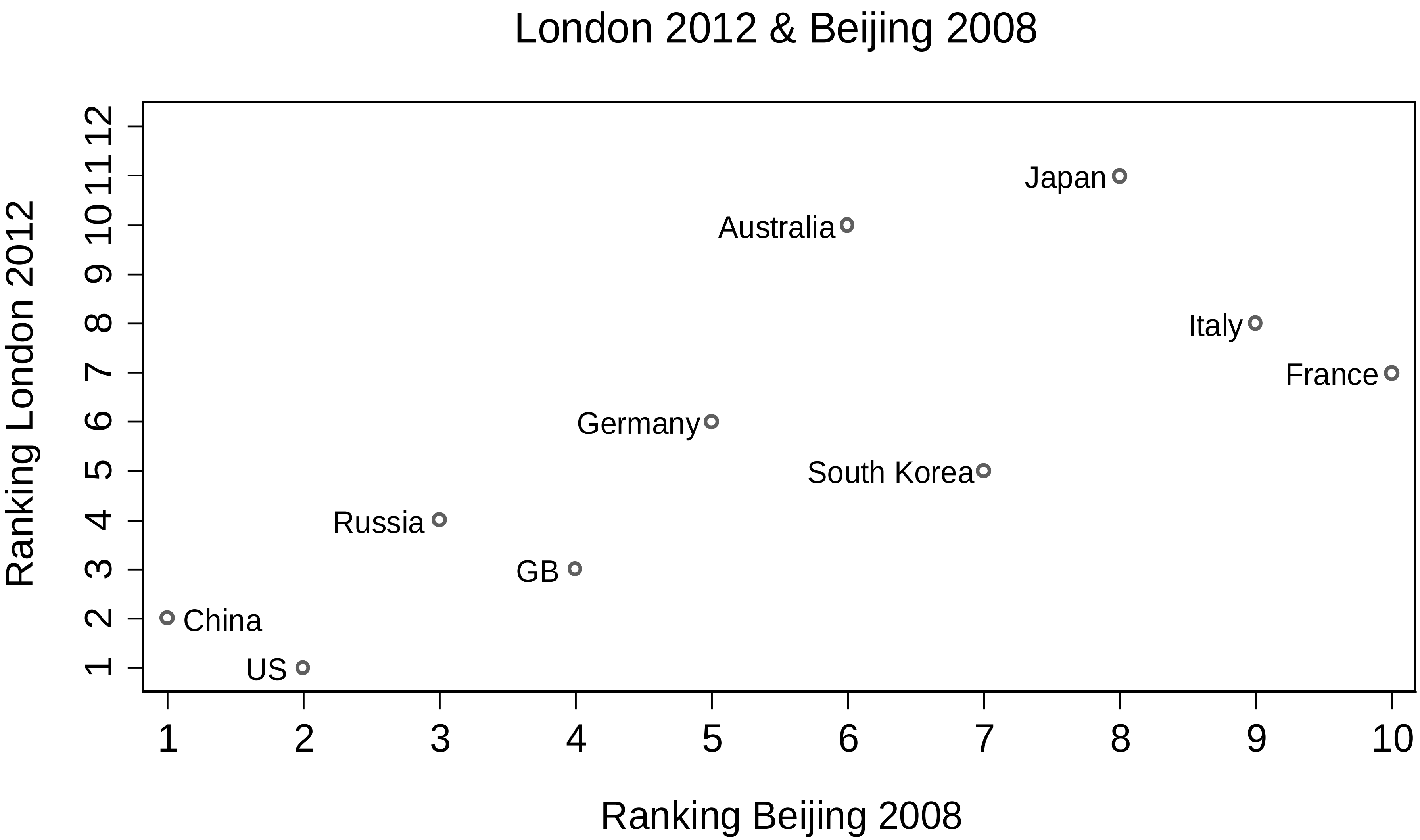
- Explain the following graphs



Olympic Gold Medals and Country Population Size

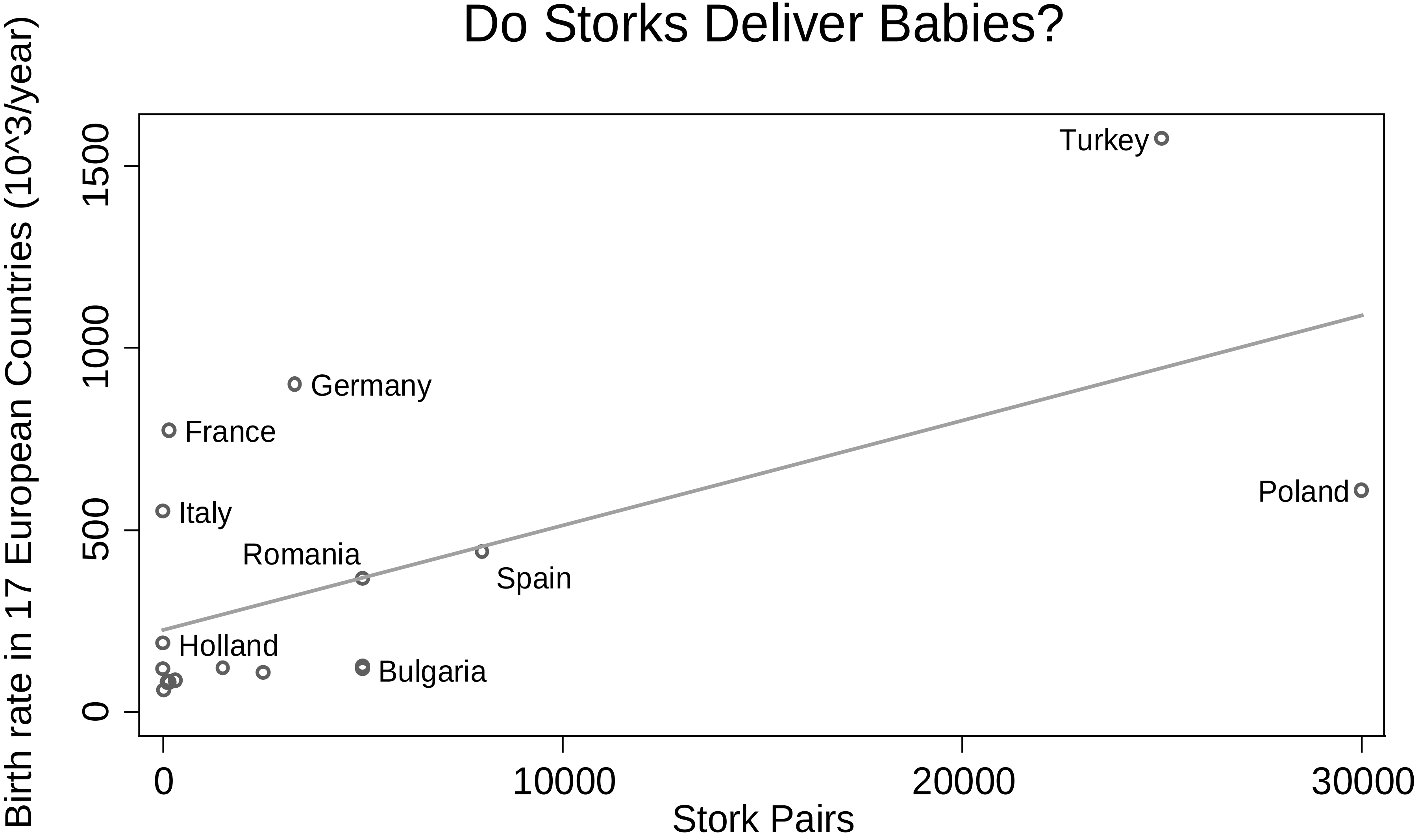
p<.01; r=.46; rsquared=.22;
Data source: Timothy Lethbridge's Blog

Olympic Medals Table Rank



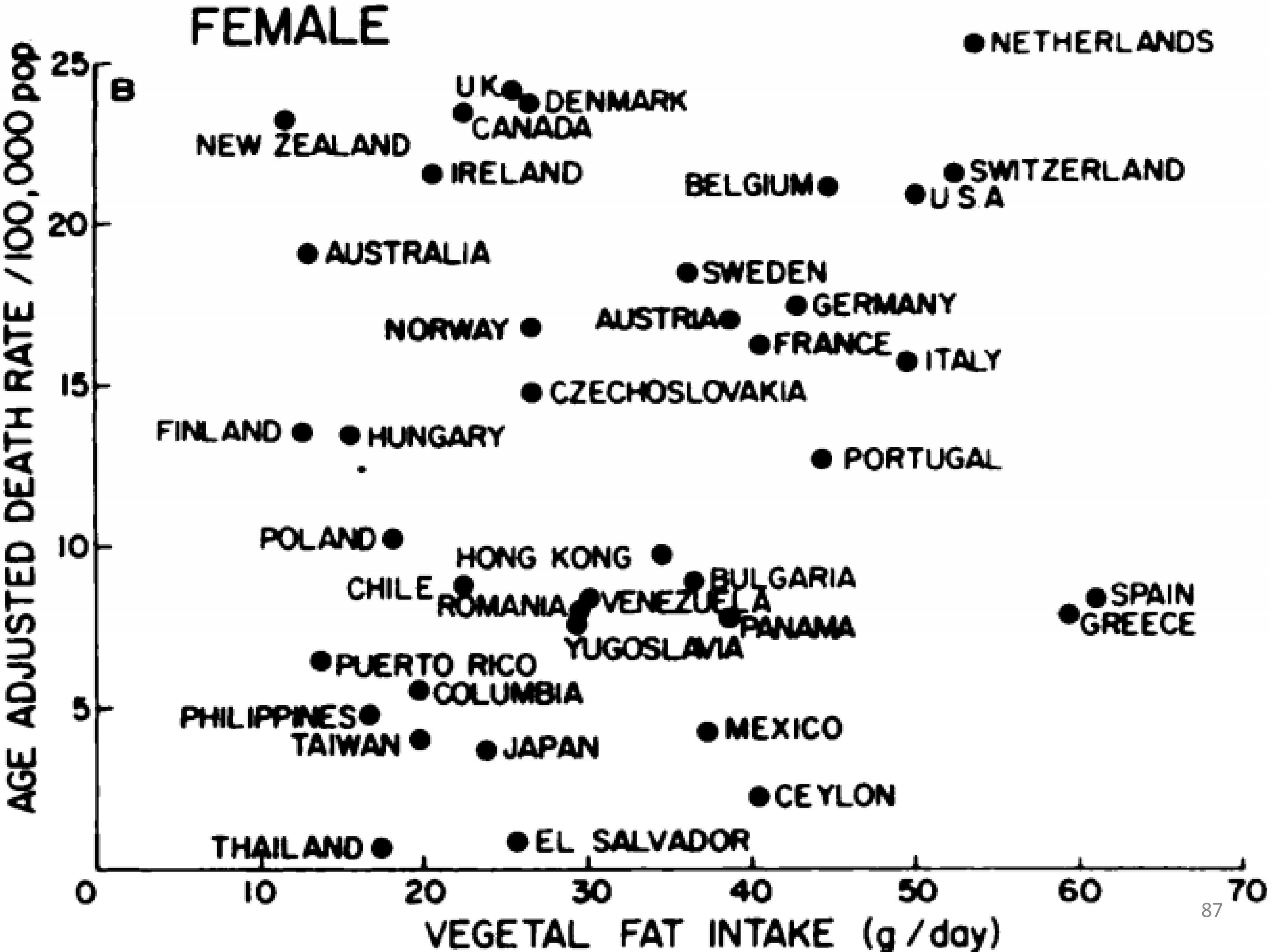
p<.01; rho=.81. Source: BBC Sport

Do Storks Deliver Babies?



p=.008; r=.62

Mathews, R. 'Storks Deliver Babies (p=0.008)'
Teaching Statistics. Volume 22, Number 2, Summer 2000

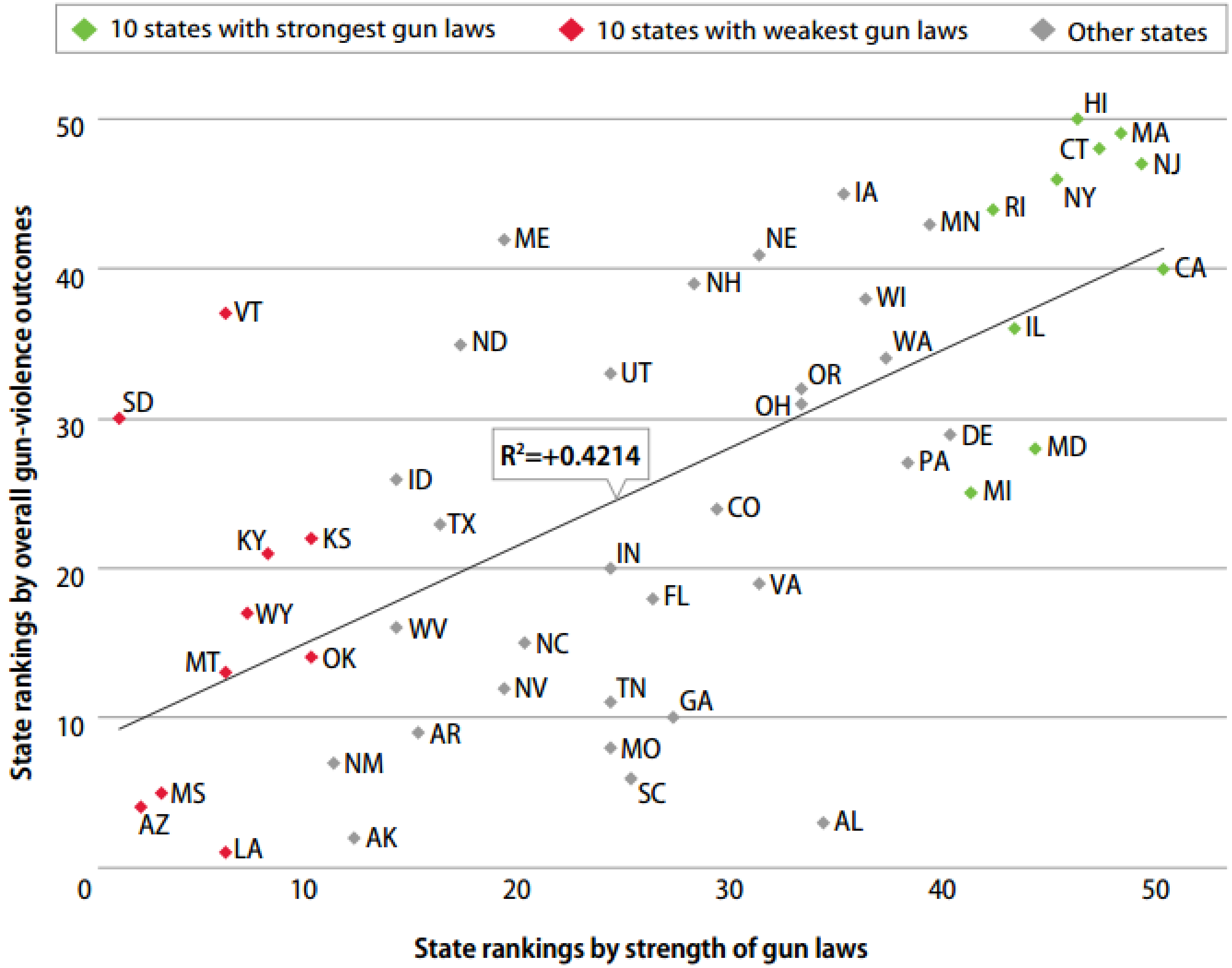

FEMALE
B
AGE ADJUSTED DEATH RATE / 100,000 pop
VEGETAL FAT INTAKE (g / day)
0
10
20
30
40
50
60
70
5
10
15
20
25
NETHERLANDS
UK
DENMARK
CANADA
NEW ZEALAND
IRELAND
SWITZERLAND
BELGIUM
U S A
AUSTRALIA
SWEDEN
GERMANY
NORWAY
AUSTRIA
FRANCE
ITALY
CZECHOSLOVAKIA
FINLAND
HUNGARY
PORTUGAL
POLAND
HONG KONG
BULGARIA
CHILE
ROMANIA
VENEZUELA
SPAIN
PANAMA
GREECE
YUGOSLAVIA
PUERTO RICO
COLUMBIA
PHILIPPINES
MEXICO
TAIWAN
JAPAN
CEYLON
THAILAND
EL SALVADOR

Activity 5

1. In pairs examine the graph below

FIGURE 3

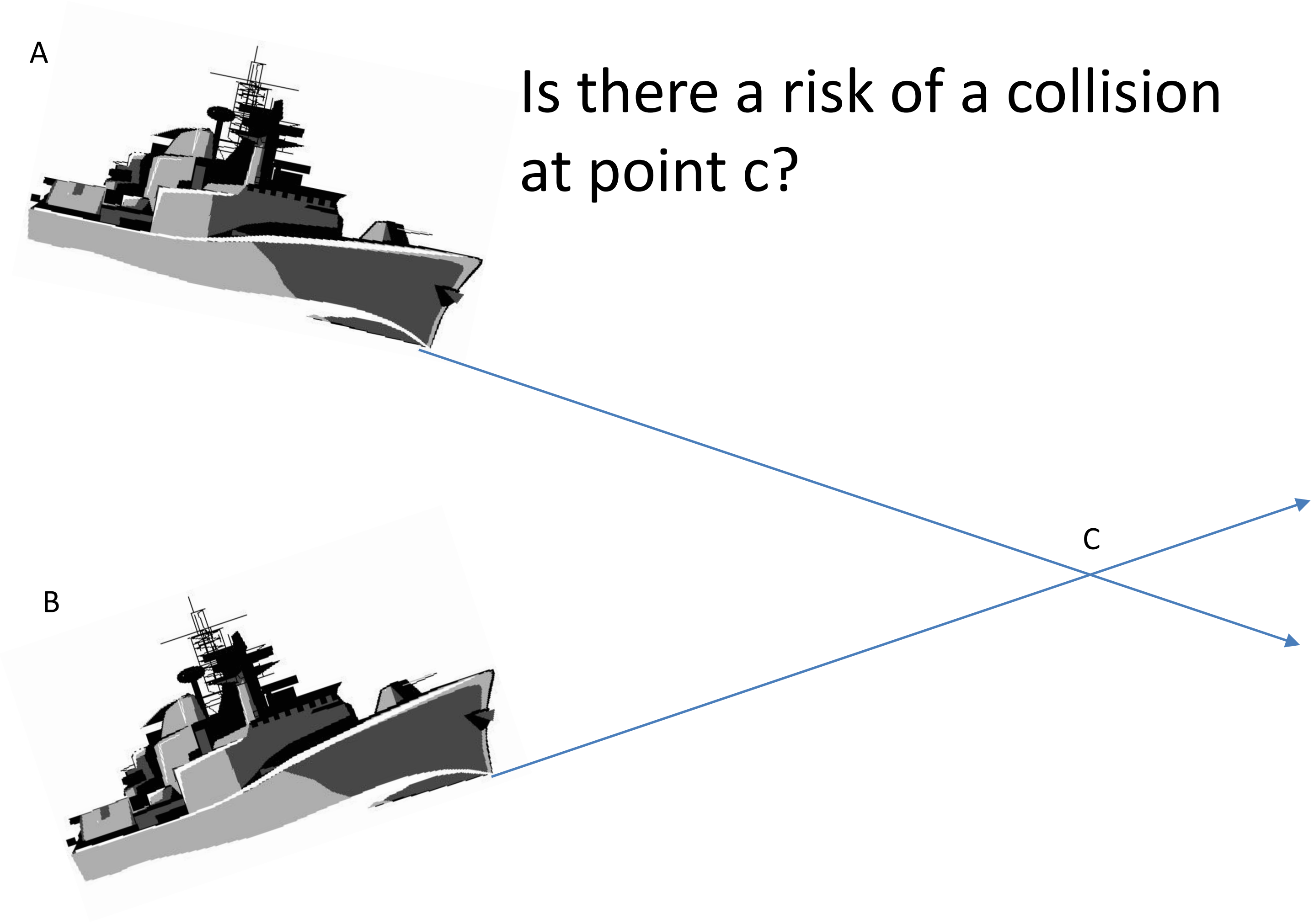
Correlation between state gun laws and gun-violence outcomes

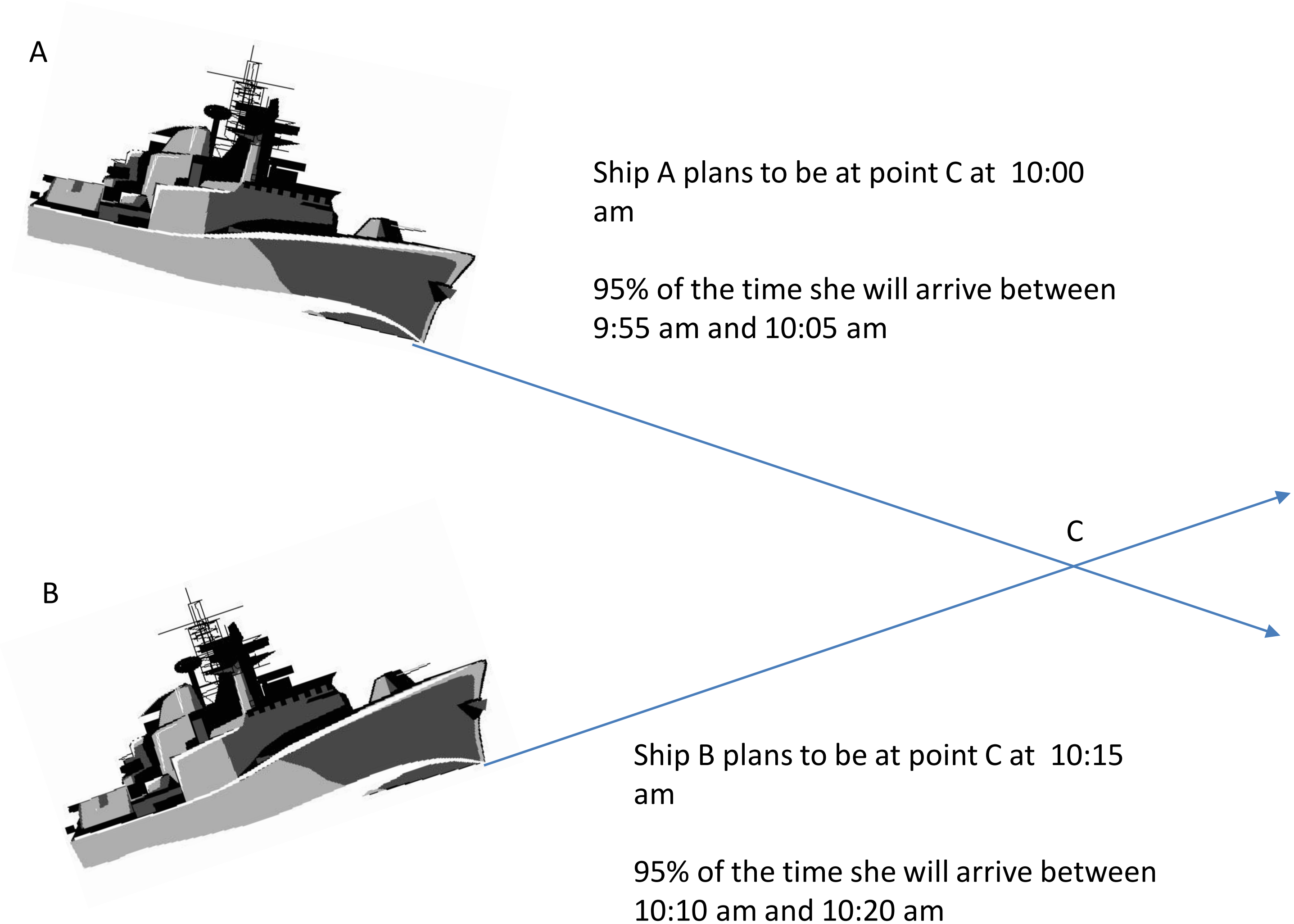


Source: Center for American Progress analysis based on data from Centers for Disease Control and Prevention, Federal Bureau of Investigation, Mayors Against Illegal Guns, and Law Center to Prevent Gun Violence.

Part 7 Confidence Intervals

Is there a risk of a collision at point c?





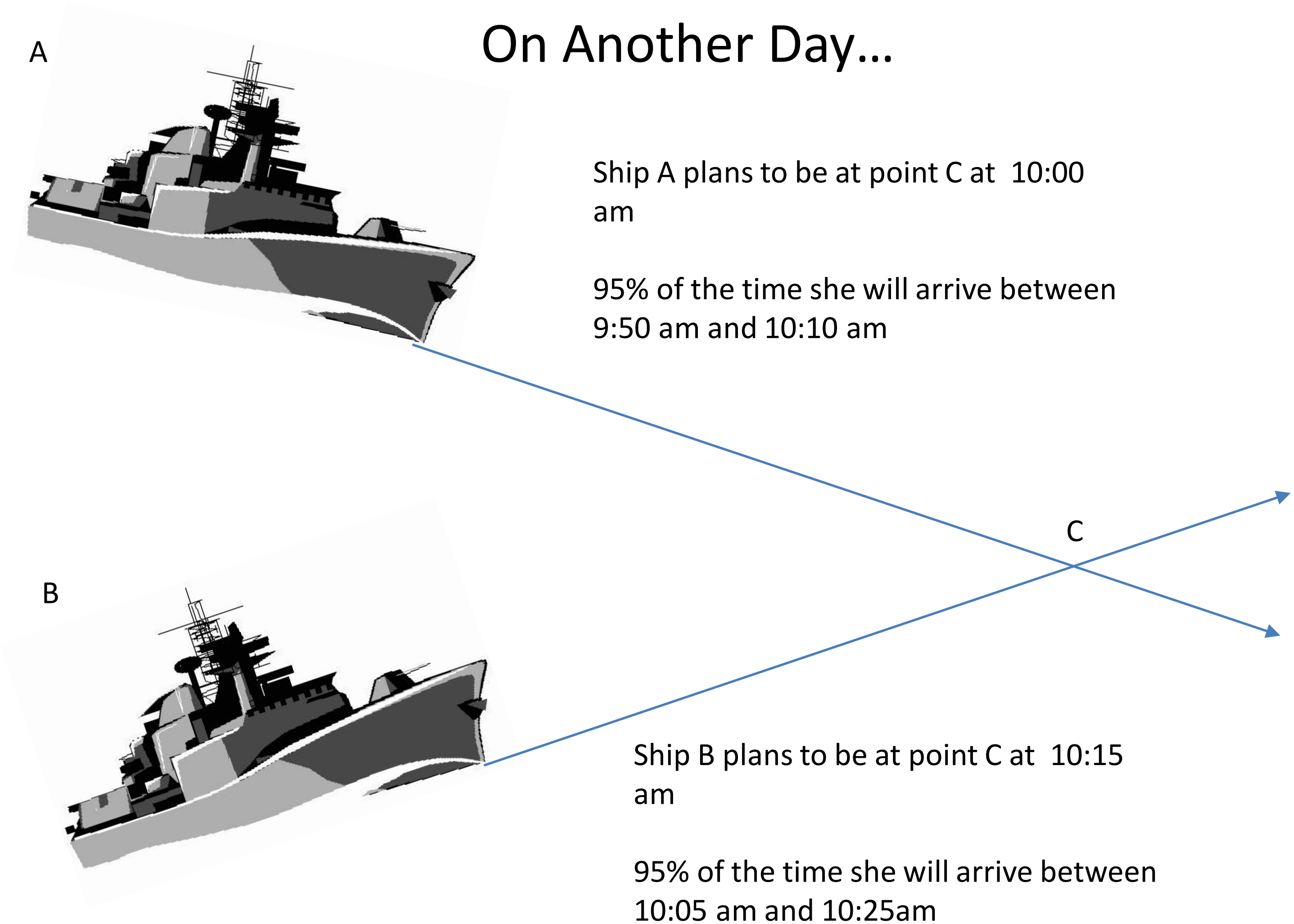
Ship A plans to be at point C at 10:00 am

95% of the time she will arrive between 9:55 am and 10:05 am

Ship B plans to be at point C at 10:15 am

95% of the time she will arrive between 10:10 am and 10:20 am

On Another Day…



Ship A plans to be at point C at 10:00 am

95% of the time she will arrive between 9:50 am and 10:10 am

Ship B plans to be at point C at 10:15 am

95% of the time she will arrive between 10:05 am and 10:25am

TAKE HOME MESSAGE

When confidence intervals overlap then the measures are not significantly different

When there is ‘clear blue water’ there is a significant difference

95% Confidence intervals around mean

$$CI = \bar{x} \pm (1.96 se_m)$$

ABSTINENCE
99.99%
EFFECTIVE

95% Confidence intervals around an estimate in a statistical model (β)

$$CI = \beta \pm (1.96 se_{\beta})$$

Professor Mac is always late…

Her wife accepts this but some times she is later than usual

Her wife keeps a diary and does some stats…

But how late is unreasonably late?

On the last ten occasions that they planned to go out she has been late nine times

- Minimum 0 minutes; Max 59 minutes
- Mean 16.4 minutes
- s.e. mean 5.4 minutes

Activity 6

1. Calculate a confidence interval

Her partner constructs a confidence interval around the mean

Upper c.i. = 16.4 + (1.96* 5.4) =	26.98
Mean	16.40
Lower c.i. = 16.4 - (1.96* 5.4) =	5.82

Another way to think about this is..

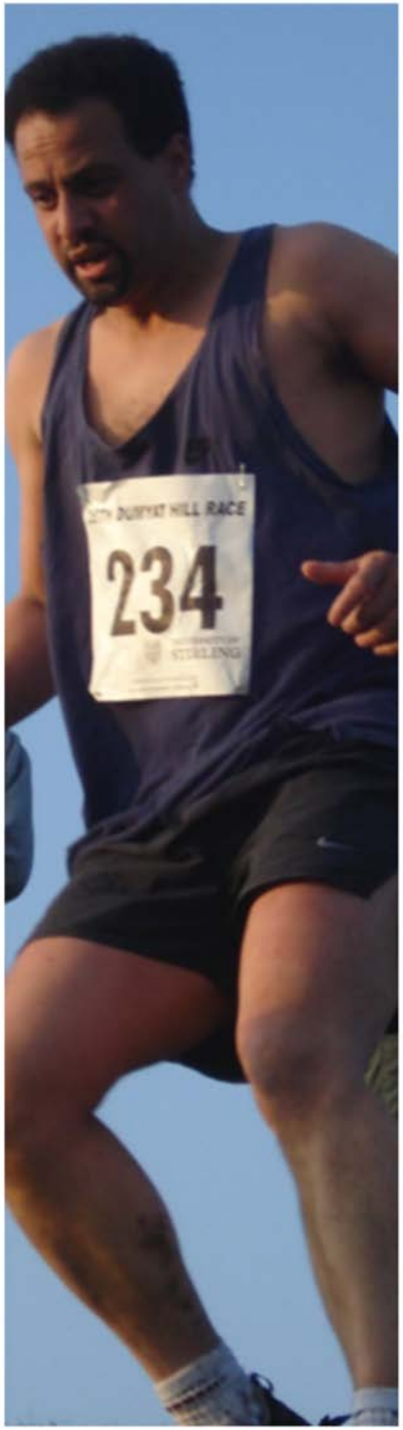
Standard error

–how tightly distributed the values are grouped around the mean

Confidence intervals

– a measure of precision of an estimate (e.g. a mean)

Part 8 Bivariate Relationships & OLS



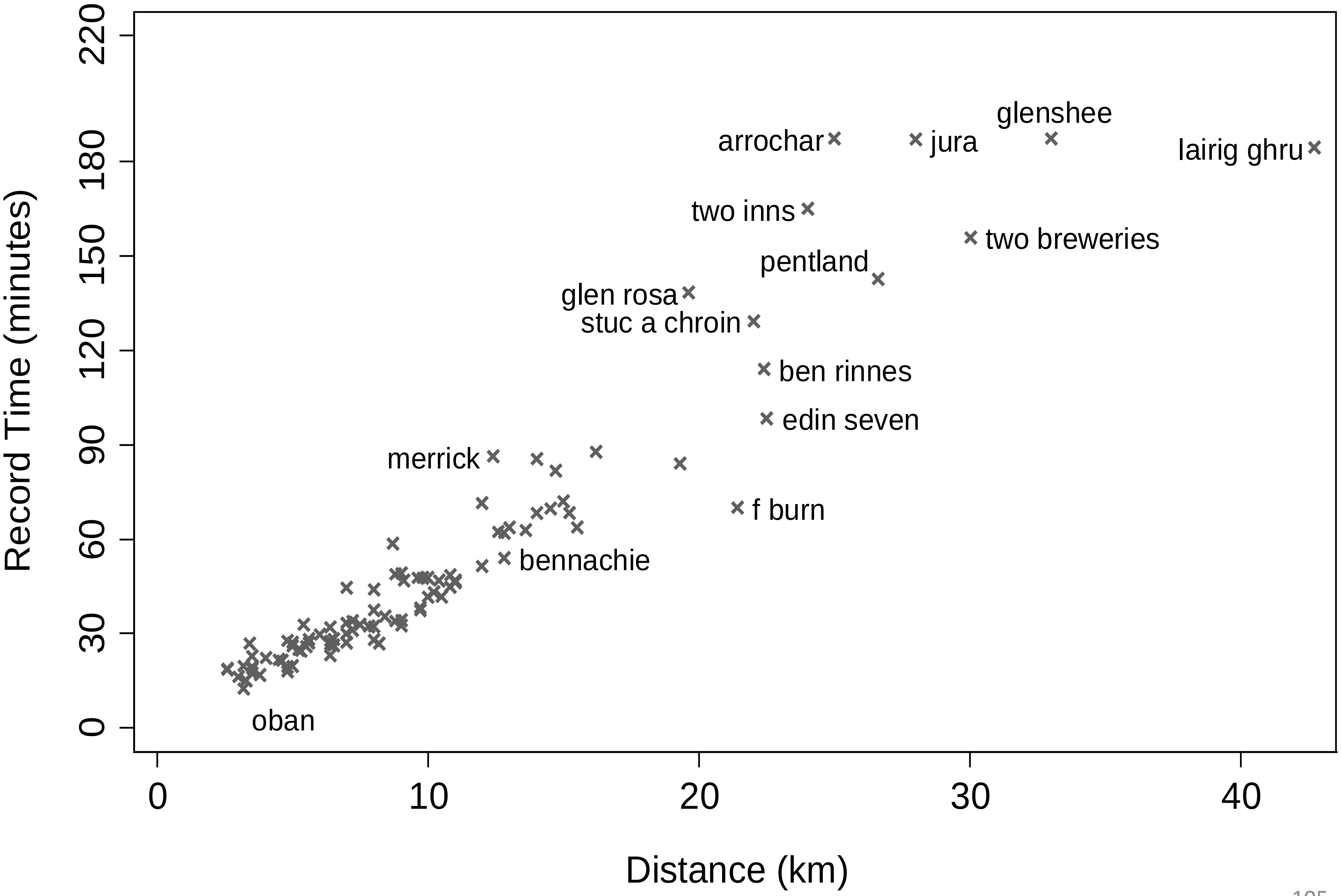
Scottish Hill Races Data

Rossie Hill Race

Angus Hunter coped well with the difficult terrain to gain first place in 34m53s ahead of Ian McArthur who finished strongly in 38m48s. Experienced cross-country runner Arabella Woodrow was first lady in a time of 40m47s followed by Alastair Bulcraig in 44m18s and Susan Lennon, 45m05s. Guest runner **Vernon Gayle** achieved 48m49s.

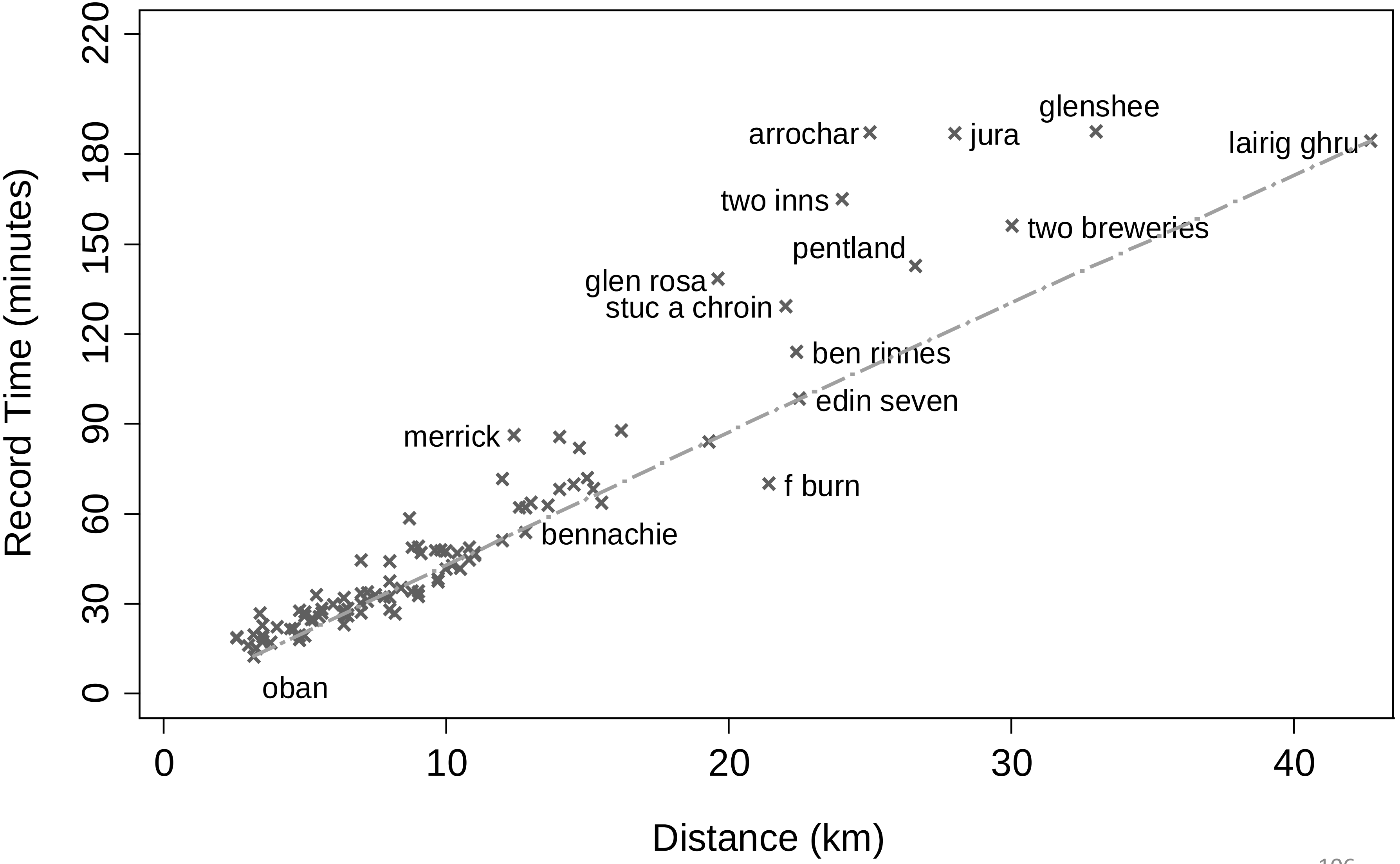
Perthshire Advertiser Friday 20th February 2004.

Scottish Hill Races - Record Times



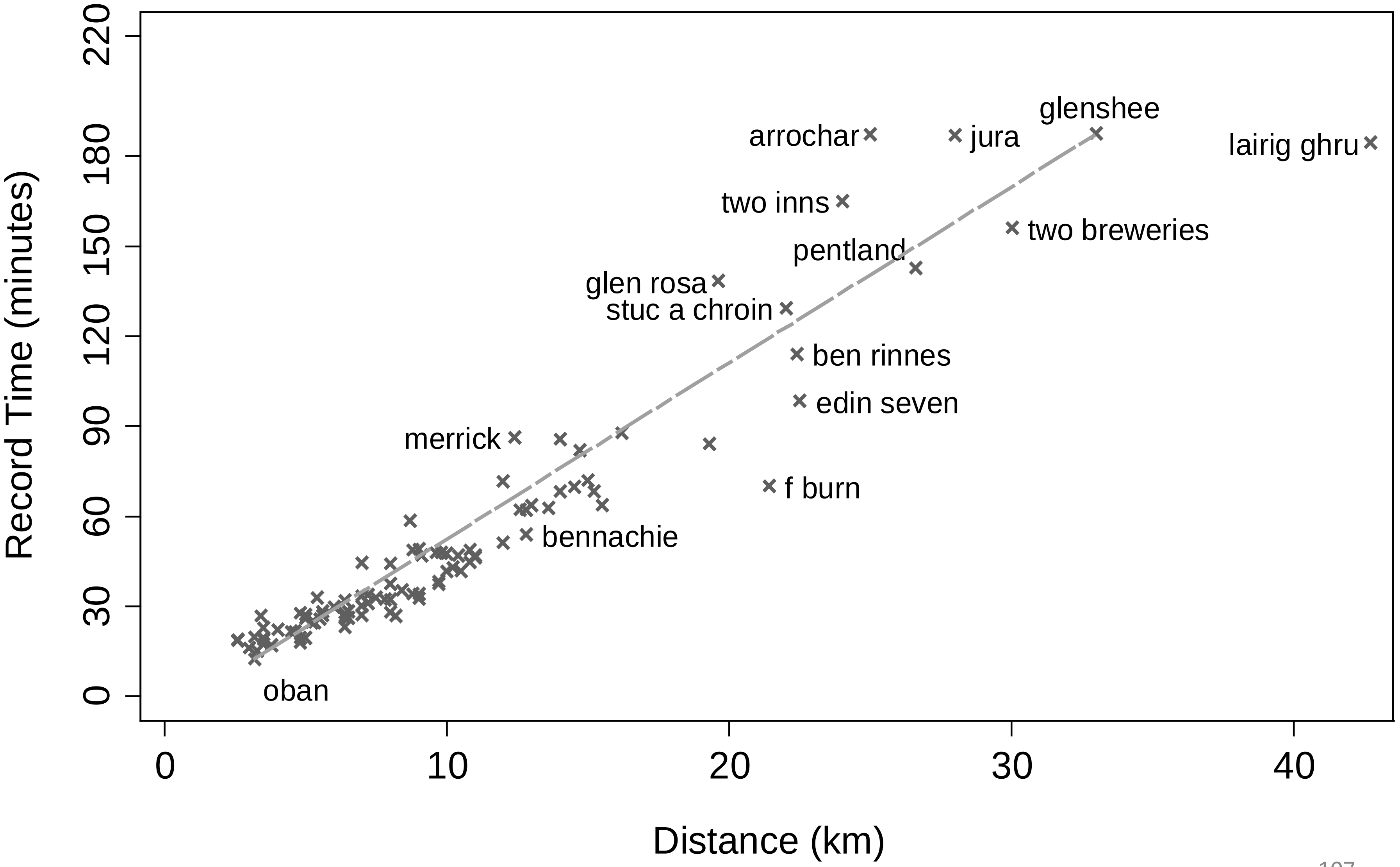
Scottish Hill Races - Record Times

(possible line of best fit)



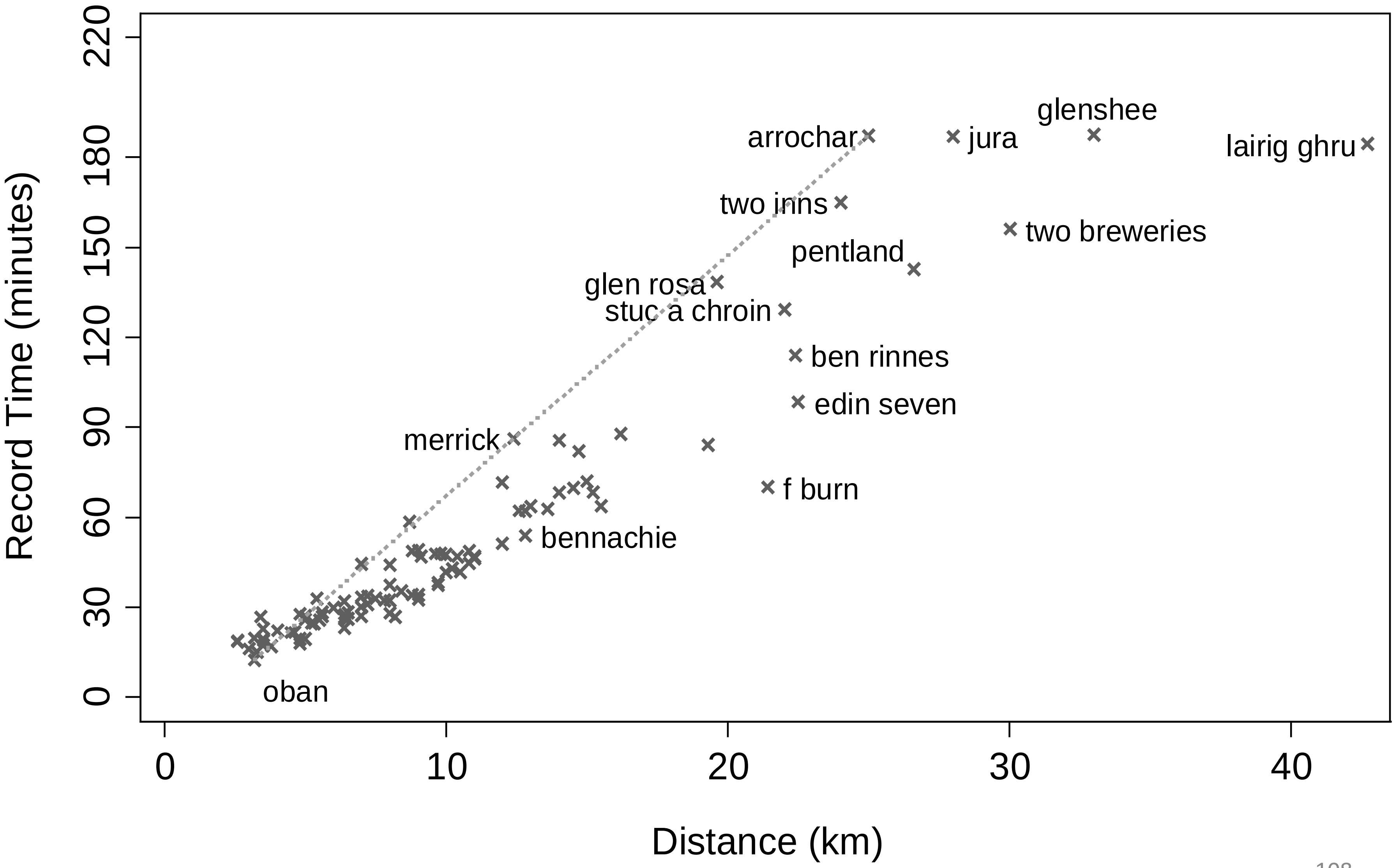
Scottish Hill Races - Record Times

(possible line of best fit)



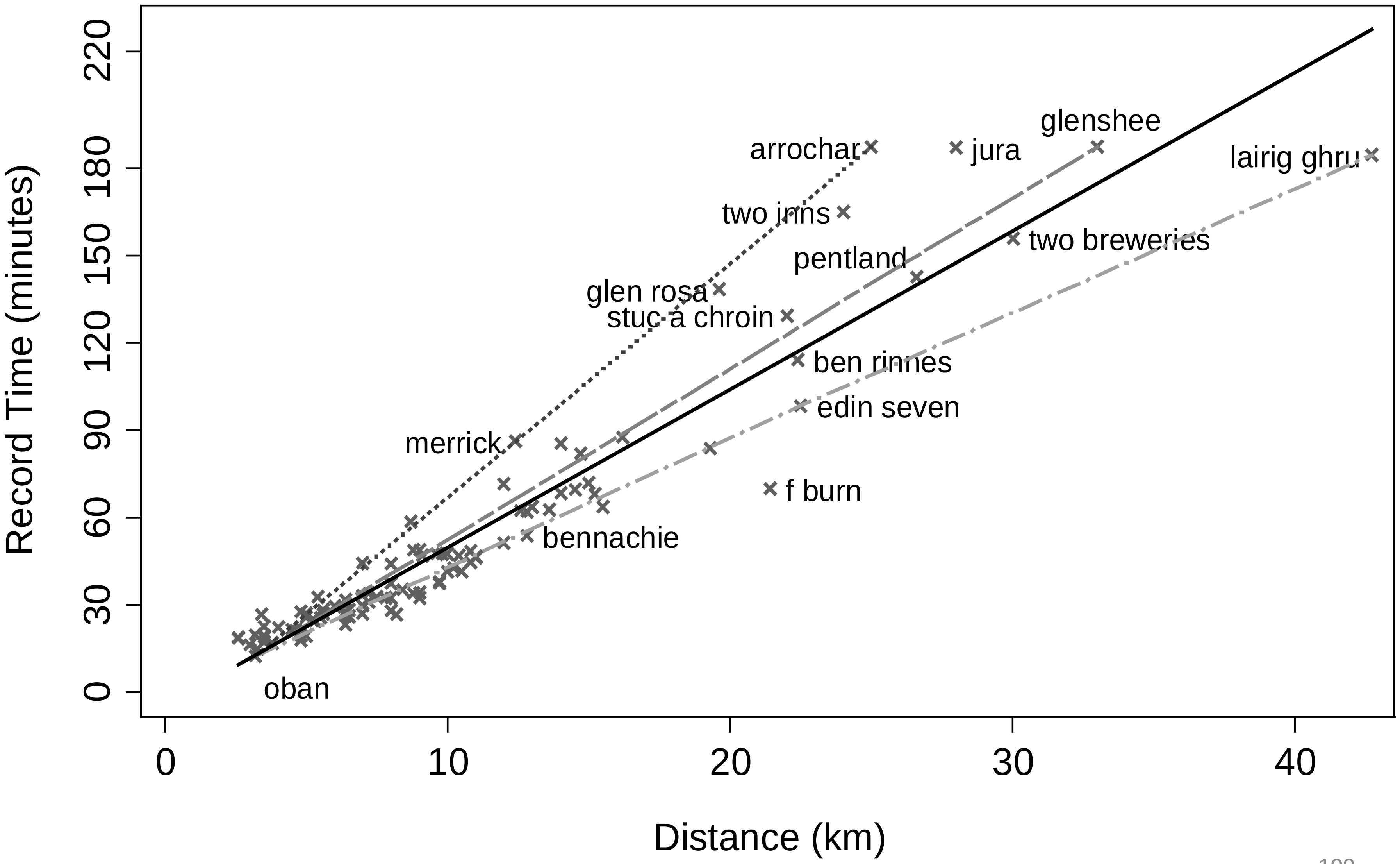
Scottish Hill Races - Record Times

(possible line of best fit)



Scottish Hill Races - Record Times

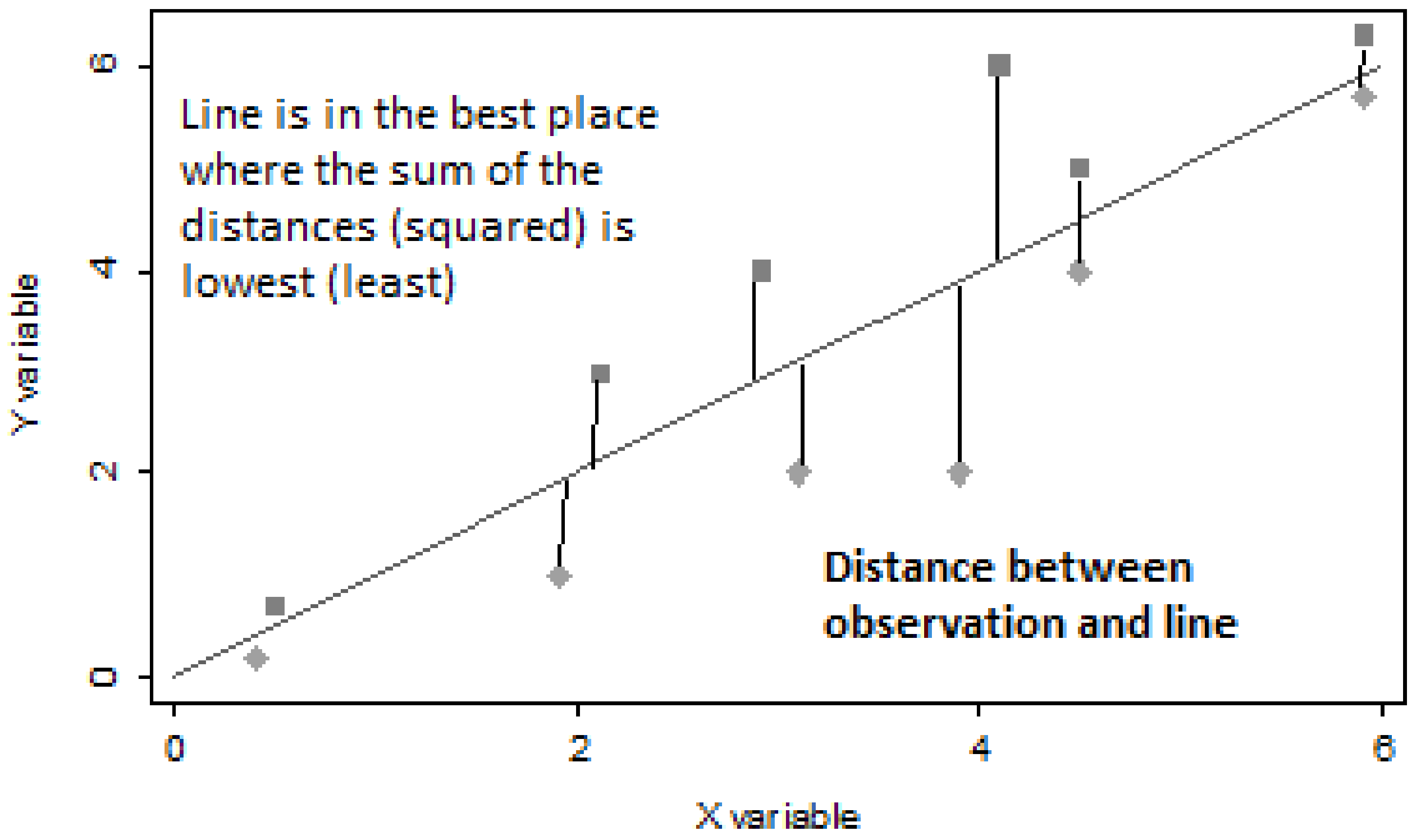
(line of best fit - solid line)



Part 9 Ordinary Least Squares

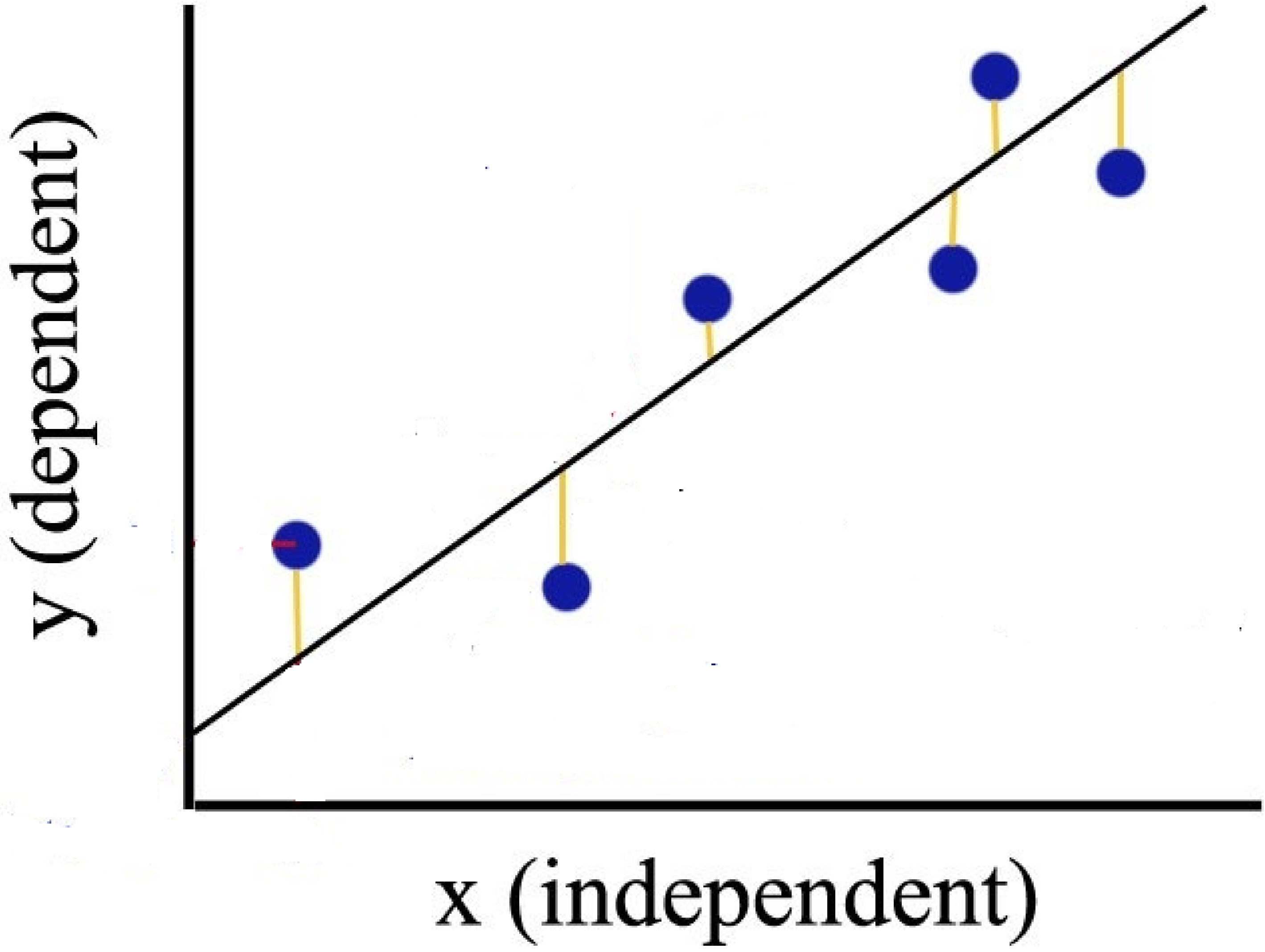
Line of Best Fit

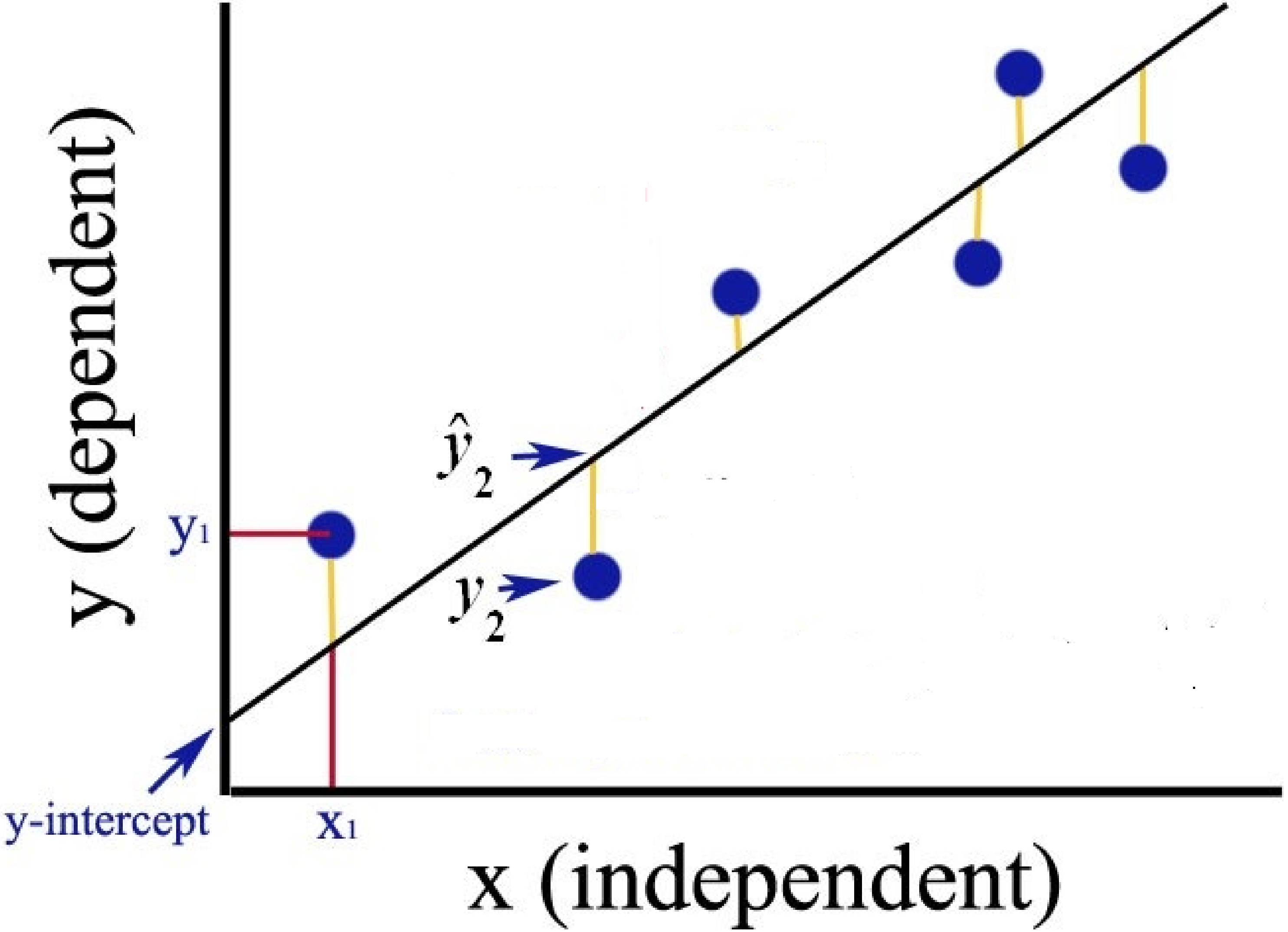
(Ordinary Least Squares)

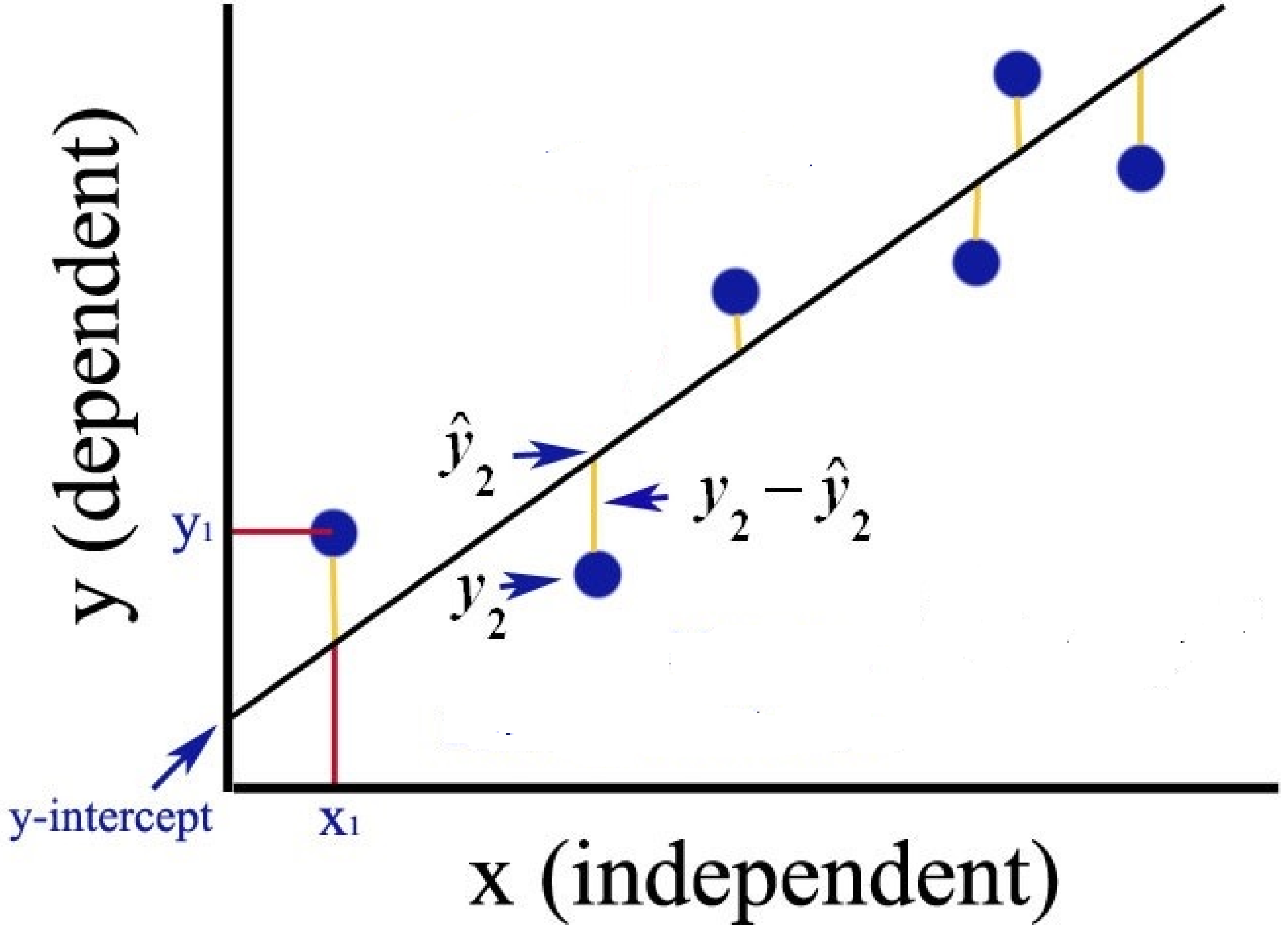


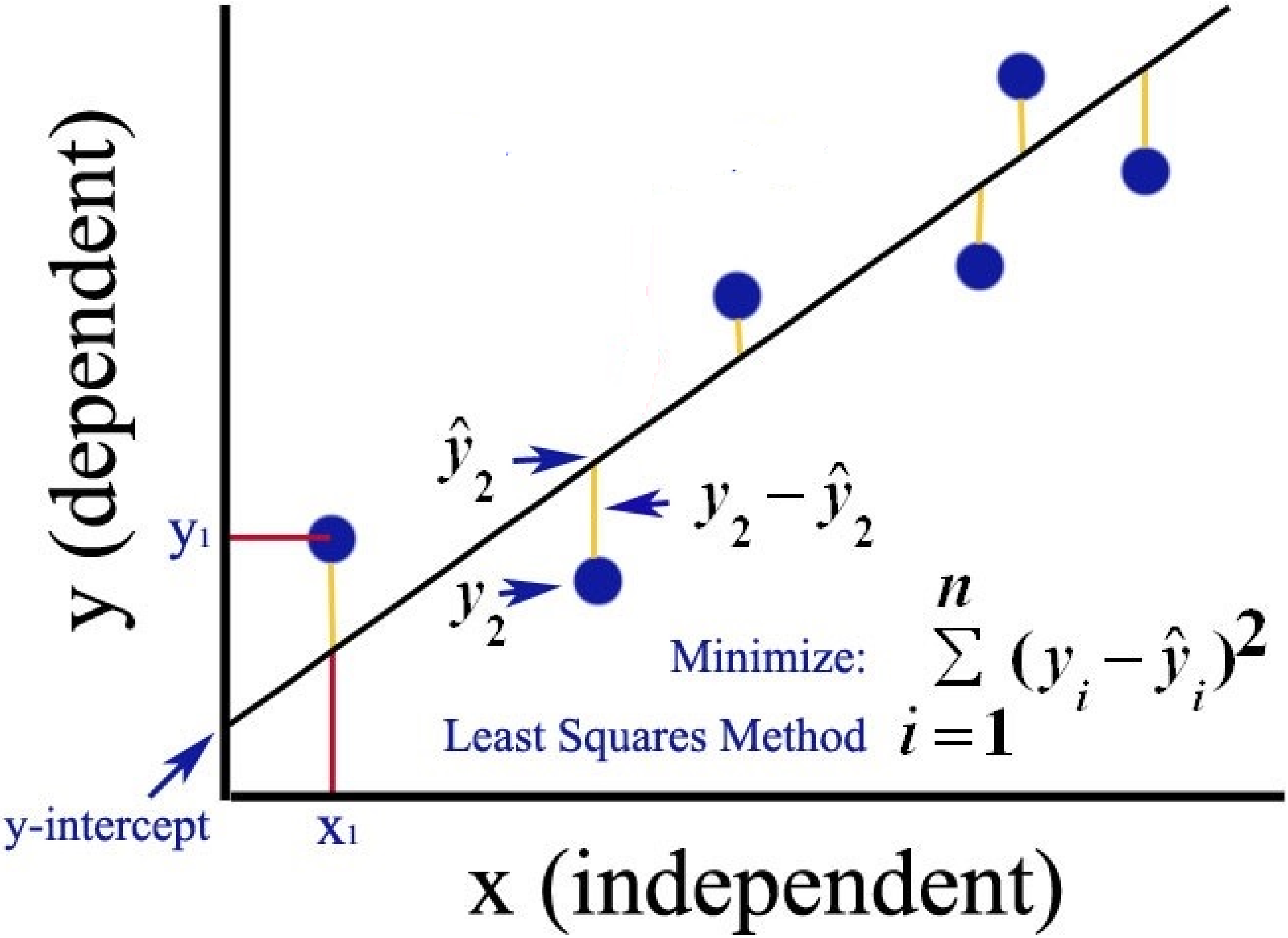
Line is in the best place where the sum of the distances (squared) is lowest (least)

Distance between observation and line


y (dependent)
x (independent)


y (dependent)
$\hat{y}_2$
y_2
y_1
y-intercept
x_1
x (independent)


y (dependent)
y₁
$\hat{y}_2$
$y_2 - \hat{y}_2$
y_2
y-intercept
x₁
x (independent)


y (dependent)
y₁
$\hat{y}_2$
$y_2 - \hat{y}_2$
y_2
Minimize: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Least Squares Method
y-intercept
x₁
x (independent)

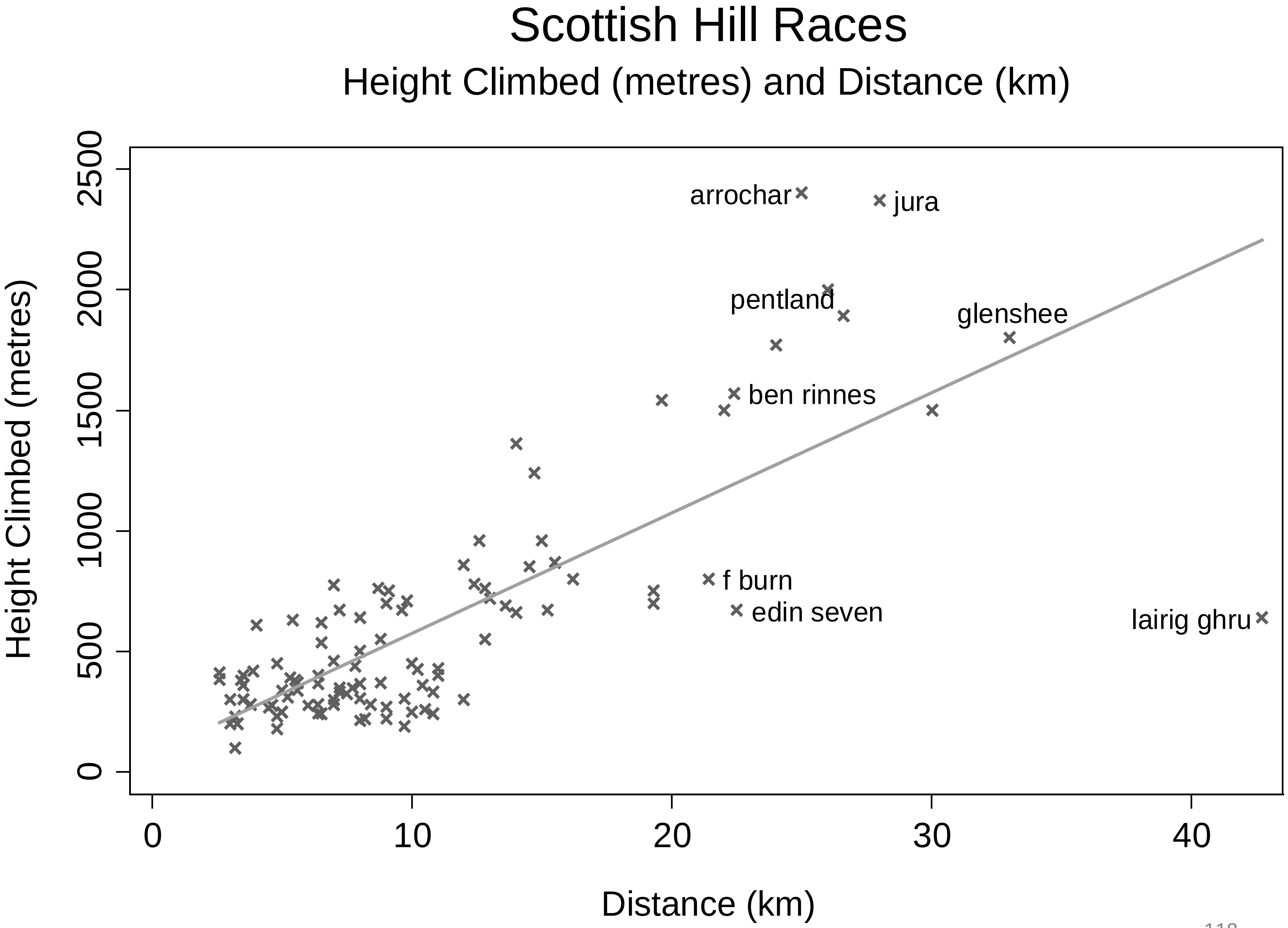
Ordinary Least Squares

- Ordinary least squares is a classical estimation method

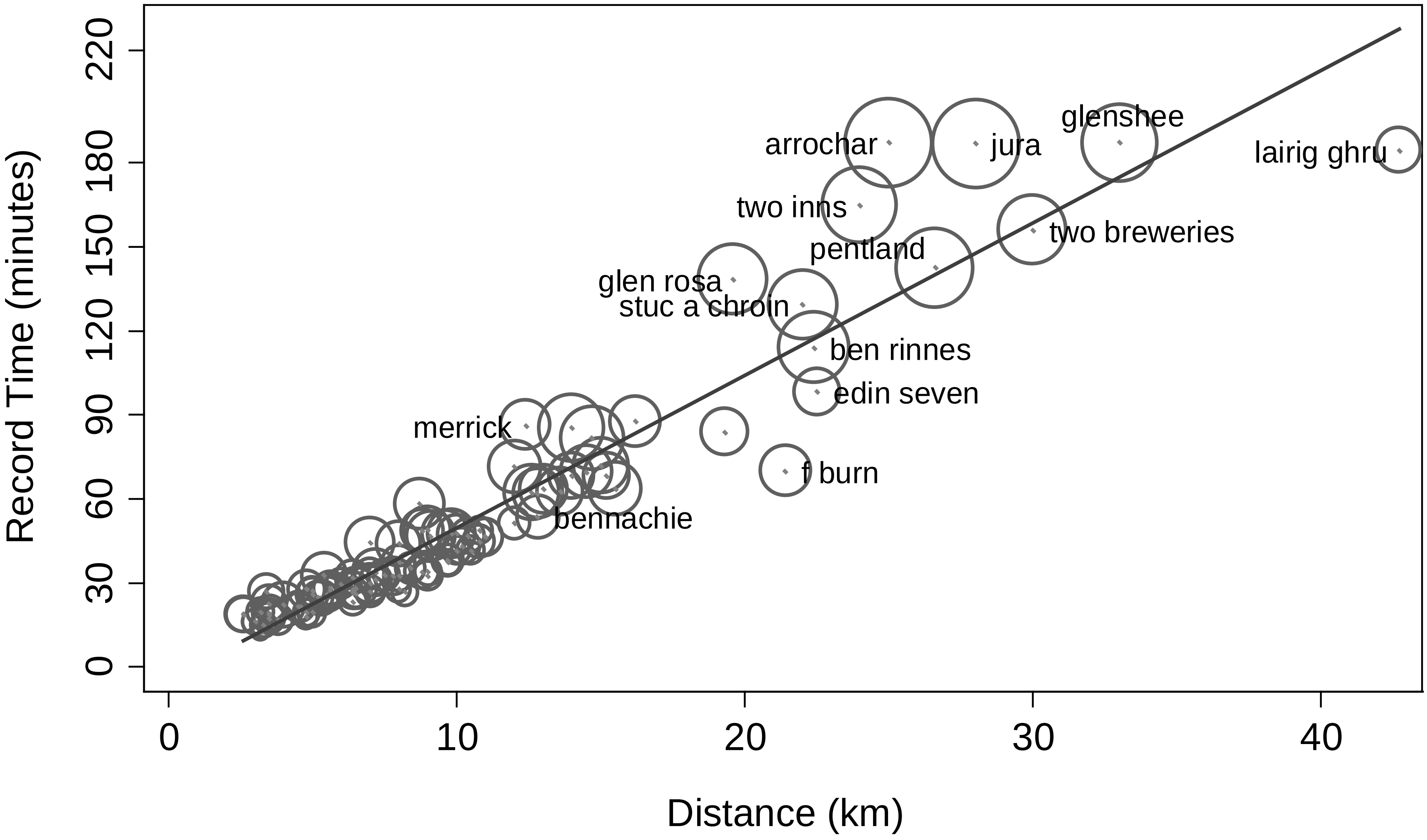
- More than two e**X**planatory variables can’t be easily be graphed
- Mathematical algorithm estimates it for many e**X**planatory variables

- Humorous Norwegians (Ordinary Least Square)

Part 10 Two Explanatory Variables

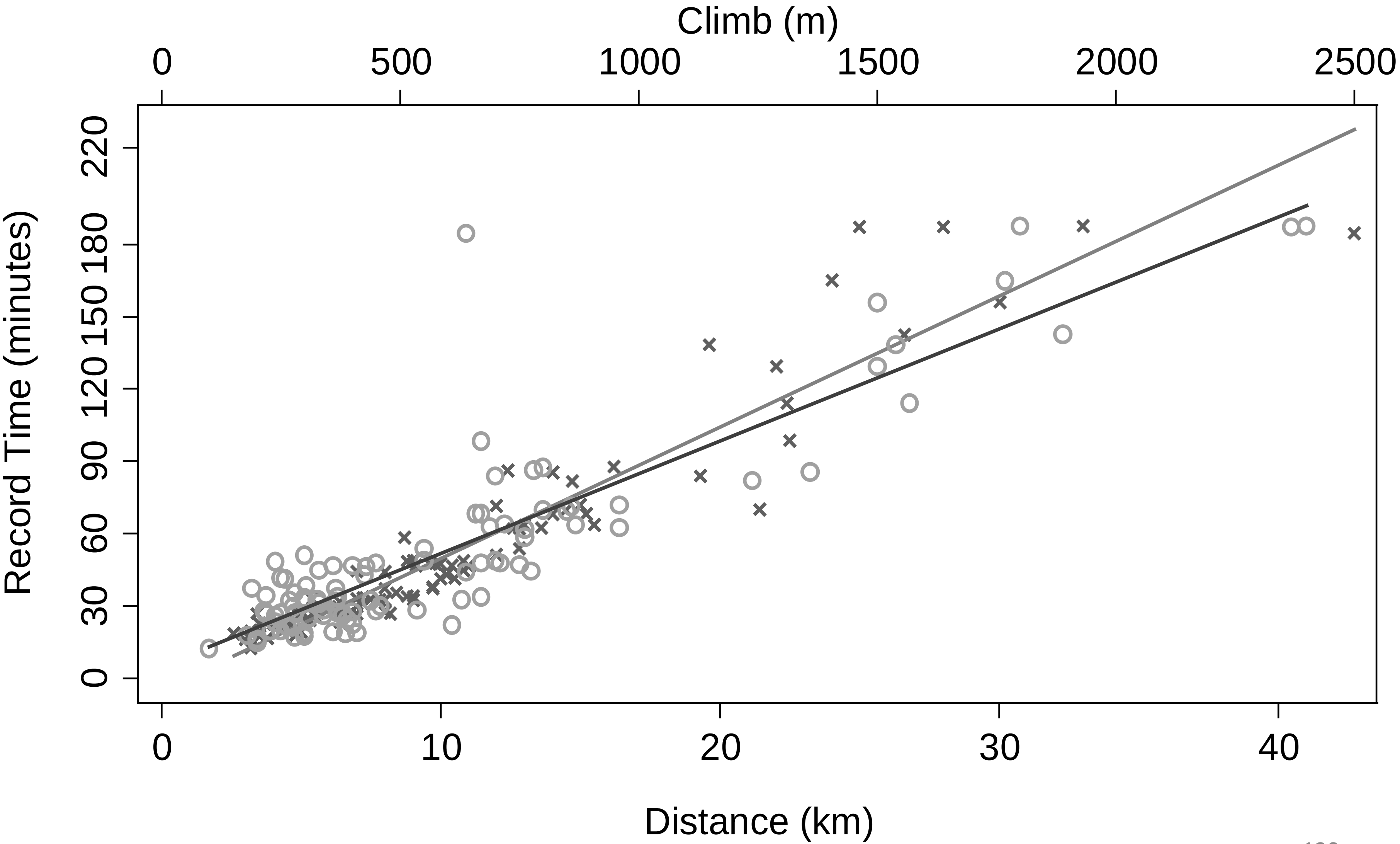

Scottish Hill Races
Height Climbed (metres) and Distance (km)
Height Climbed (metres)
2500
2000
1500
1000
500
0
0
10
20
30
40
Distance (km)
arrochar
jura
pentland
glenshee
ben rinnes
f burn
edin seven
lairig ghru

Scottish Hill Races - Record Times



(markers weighted - metres climbed)

Scottish Hill Races - Record Times



Part 11 Statistical Models (part 1)

Statistical models augment our ability to investigate this complicated social world

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”

Take Home Message #3

A standard regression model has ONE OUTCOME variable

and

MULTIPLE EXPLANATORY variables

Take Home Message #4

A statistical model will usually tell you two things

1. Which variables are important (i.e. significant)

2. The effect of the important variables (i.e. strength)

Four Things to Remember

1. (Adjusted) R^2

2. p values for each e**X**planatory variable

3. Beta (β) sign for each e**X**planatory variable

4. Beta (β) size for each e**X**planatory variable

Things to Remember # 1

- R^2 the Coefficient of Multiple Determination

  Takes on values between 0 and 1

  In a good journal Adjusted R^2 will be reported

- 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

Things to Remember # 2

- **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’

- Interpret the **p value** in the usual way (i.e. is it less than .05 etc)

Things to Remember # 3

- 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)

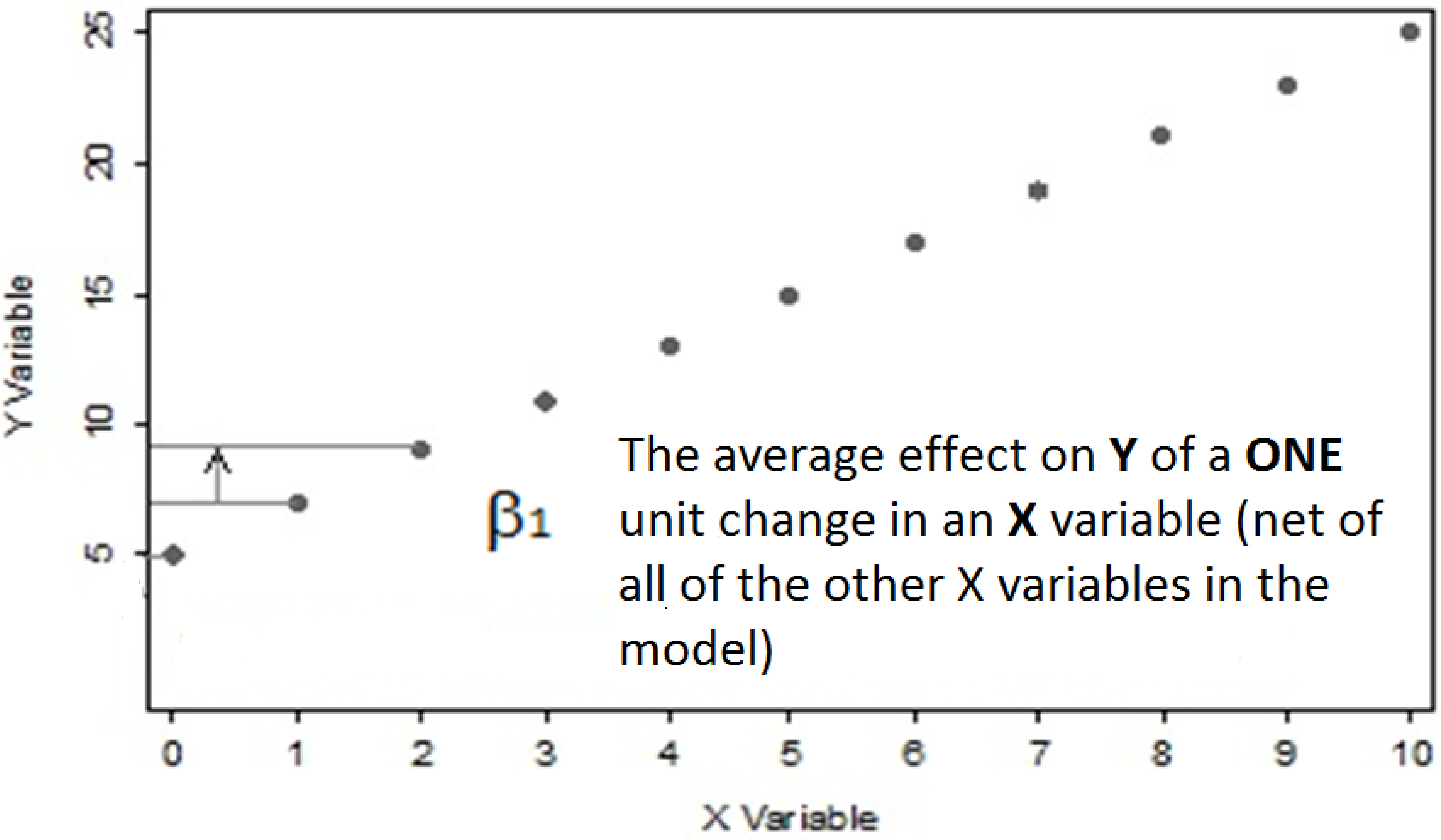
Things to Remember # 4

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)

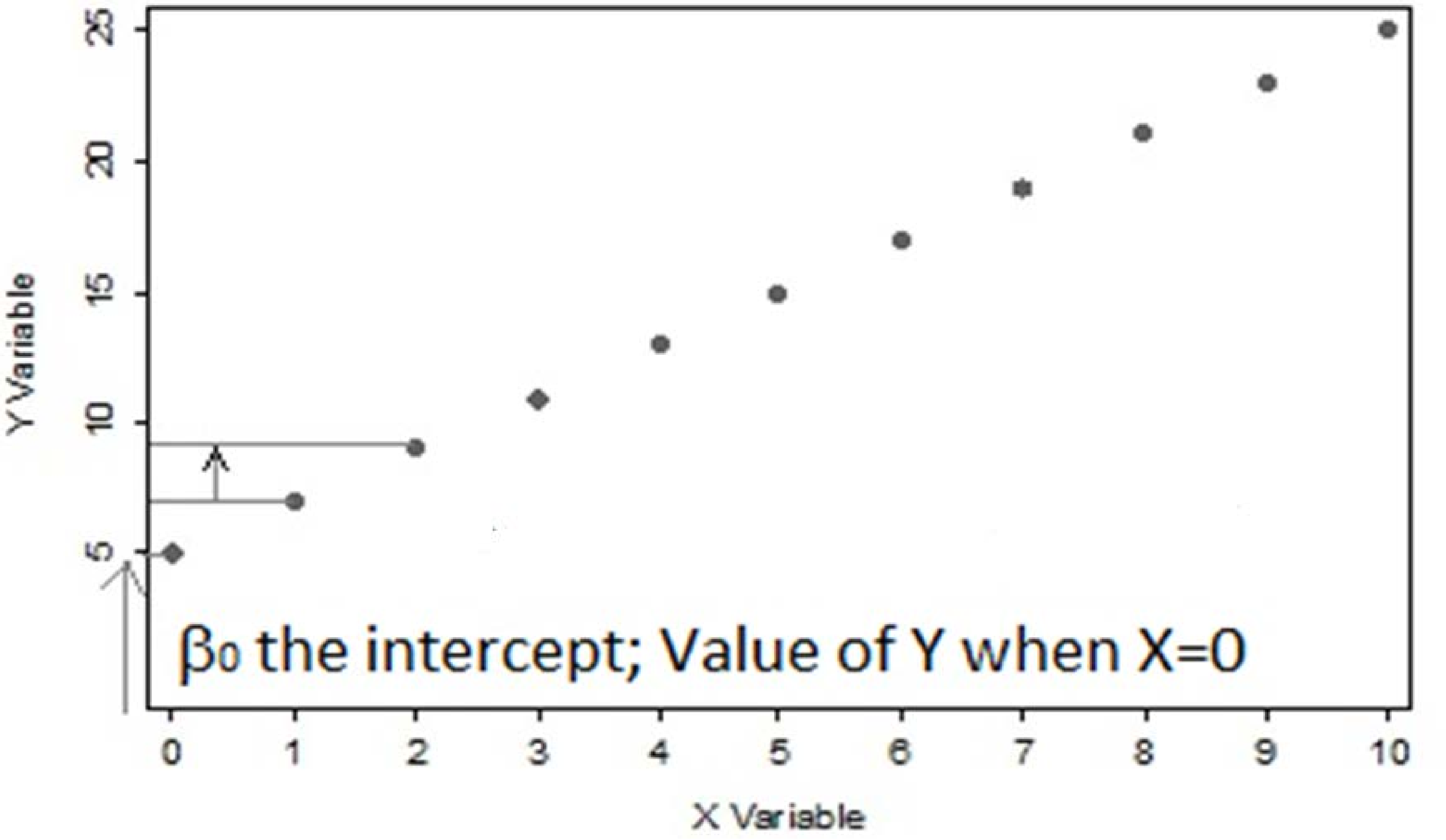
(In some example (β) is standardized so all e**X**planatory variables are on the same scale)

Things to Remember # 4 *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)

Could Potentially Remember…
The Constant is Beta Zero (β_0)



β_0 the intercept; Value of Y when X=0

Four Things to Remember

1. (Adjusted) R^2

2. p values for each e**X**planatory variable

3. Beta (β) sign for each e**X**planatory variable

4. Beta (β) size for each e**X**planatory variable

Part 12 More Statistical Models

Example 1

Modelling the Hill Racing Records

Y variable Record (in minutes)

e**X**planatory variables

X_1 Distance (km)

X_2 Climb (metres)

Modelling the Hill Racing Records

Adjusted $R^2 = .97$

Variable	*p value*	*Beta*
Distance (1km)	<.01	3.68 (extra minutes)
Climb (1m)	<.01	0.04 (about 2 seconds)

Mo Farah 10km Gold in 2012 was apprx 2.8 minutes per 1km

Example 2 Conditions of fertility decline in developing countries, 1965-75

Change in crude birth rate (CBR) between 1965 and 1975, for 20 countries in Latin America

Reference: P.W. Mauldin and B. Berelson (1978) Conditions of fertility decline in developing countries, 1965-75, *Studies in Family Planning,9:89-147*

JSTOR: http://www.jstor.org/stable/1965523

Y variable percentage decline in the crude birth rate

e**X**planatory variables

$\mathbf{X_1}$ index of family planning effort

$\mathbf{X_2}$ index of social setting

Simple Model of Fertility Decline

Adjusted $R^2 = .62$

Variable	*p value*	*Beta*
FP Effort	<.001	1.3 (% decline in fertility)

Model of Fertility Decline

Adjusted R^2 = .71 (increase of .09)

Variable	*p value*	*Beta*
FP Effort	<.001	1.3 (% decline in fertility)
Setting	.02	0.3 (% decline in fertility)

Example 3 High School and Beyond

US Dataset (1980) n=200 students

Y Science test score

e**X**planatory variables

$\mathbf{X}_1$ Maths test score
$\mathbf{X}_2$ Reading score
$\mathbf{X}_3$ Writing score
$\mathbf{X}_4$ Social science score
$\mathbf{X}_5$ Private school

Model of Science Test Score

Adjusted R^2 = .52

Variable	*p value*	*Beta*
Maths	<.01	.32
Reading	<.01	.31
Writing	<.01	.22
Social Sci	.73	-.02 (not significant = 0)
Private Sch	.74	-.45 (compared with Public)

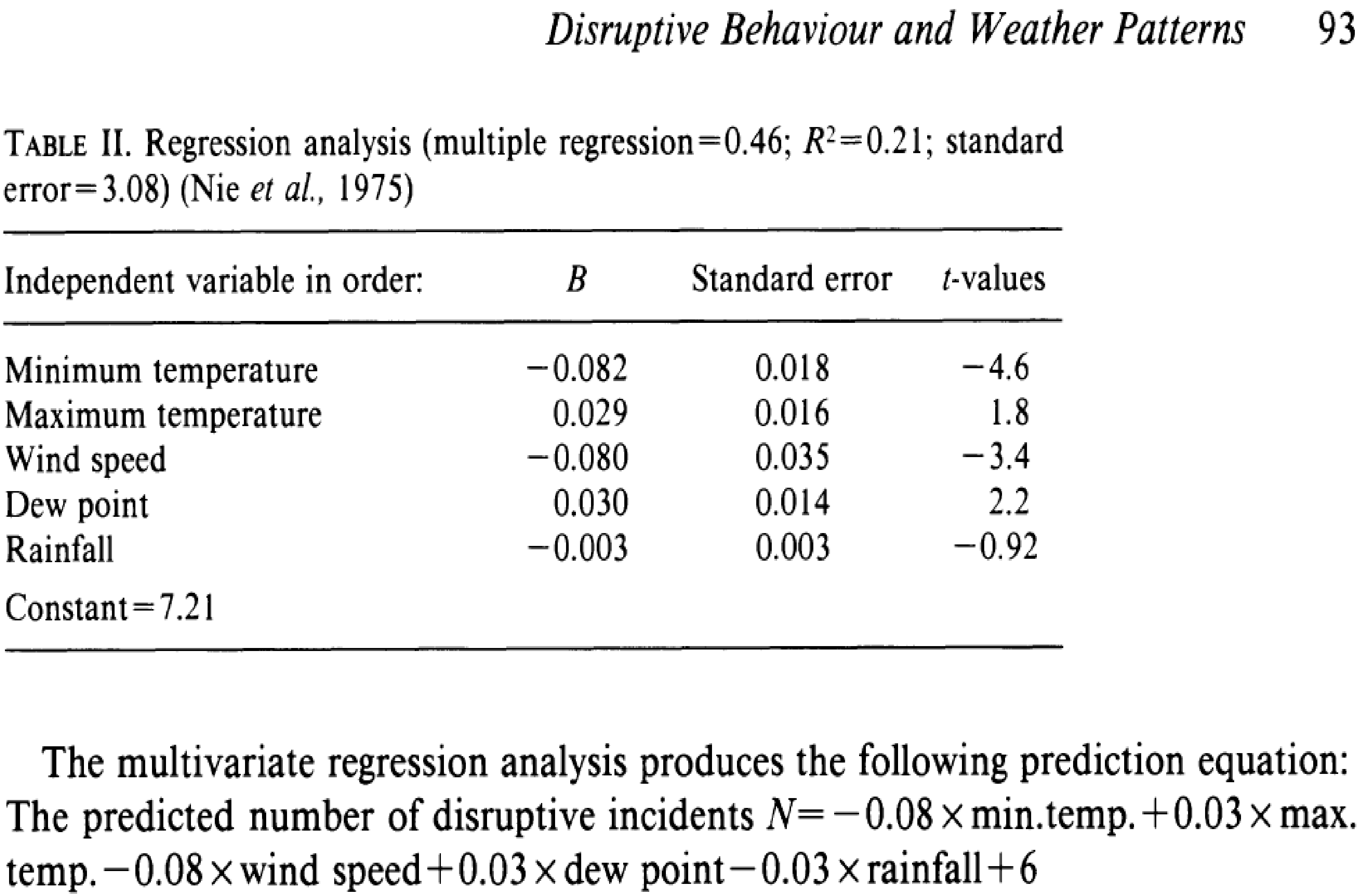


TABLE II. Regression analysis (multiple regression=0.46; $R^2=0.21$; standard error=3.08) (Nie *et al.*, 1975)

Independent variable in order:	*B*	Standard error	*t*-values
Minimum temperature	−0.082	0.018	−4.6
Maximum temperature	0.029	0.016	1.8
Wind speed	−0.080	0.035	−3.4
Dew point	0.030	0.014	2.2
Rainfall	−0.003	0.003	−0.92
Constant=7.21			

The multivariate regression analysis produces the following prediction equation: The predicted number of disruptive incidents $N=-0.08\times\text{min.temp.}+0.03\times\text{max. temp.}-0.08\times\text{wind speed}+0.03\times\text{dew point}-0.03\times\text{rainfall}+6$

Part 13 Statistical Software

- **Stata** The greatest statistical software
- **SPSS** A good enough package
- **SAS** Govt use it, can be fiddly and takes effort to learn
- **Minitab** Now a little less common but still suitable

- **R** A programming language, harder to learn but it is free
- **Python** Data science language, hard to learn but it is free

Part 14 Statistical Models in Research

Other Types of Statistical Model

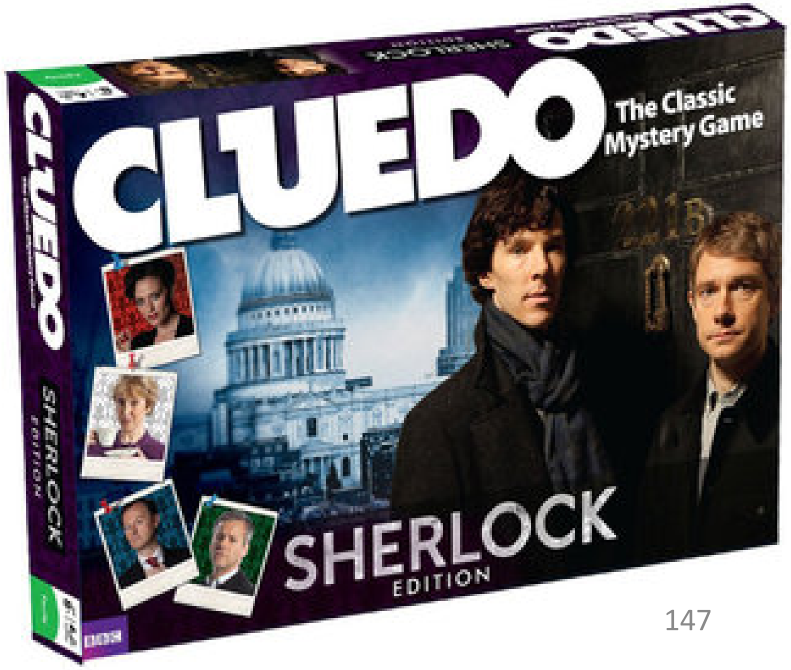
Outcome	**Model**
Binary Y	Logit
Binary Y	Probit
Categorical Y	Multinomial logit
Count Y	Poisson
Ordered categories Y	Proportional odds
Ordered categories Y	Continuation ratio

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



Why More Complex Data Analysis?

My view (although it might be controversial)…

In reality it is unlikely that a bivariate (two explanatory variable) explanation will capture the complexity of the real social world

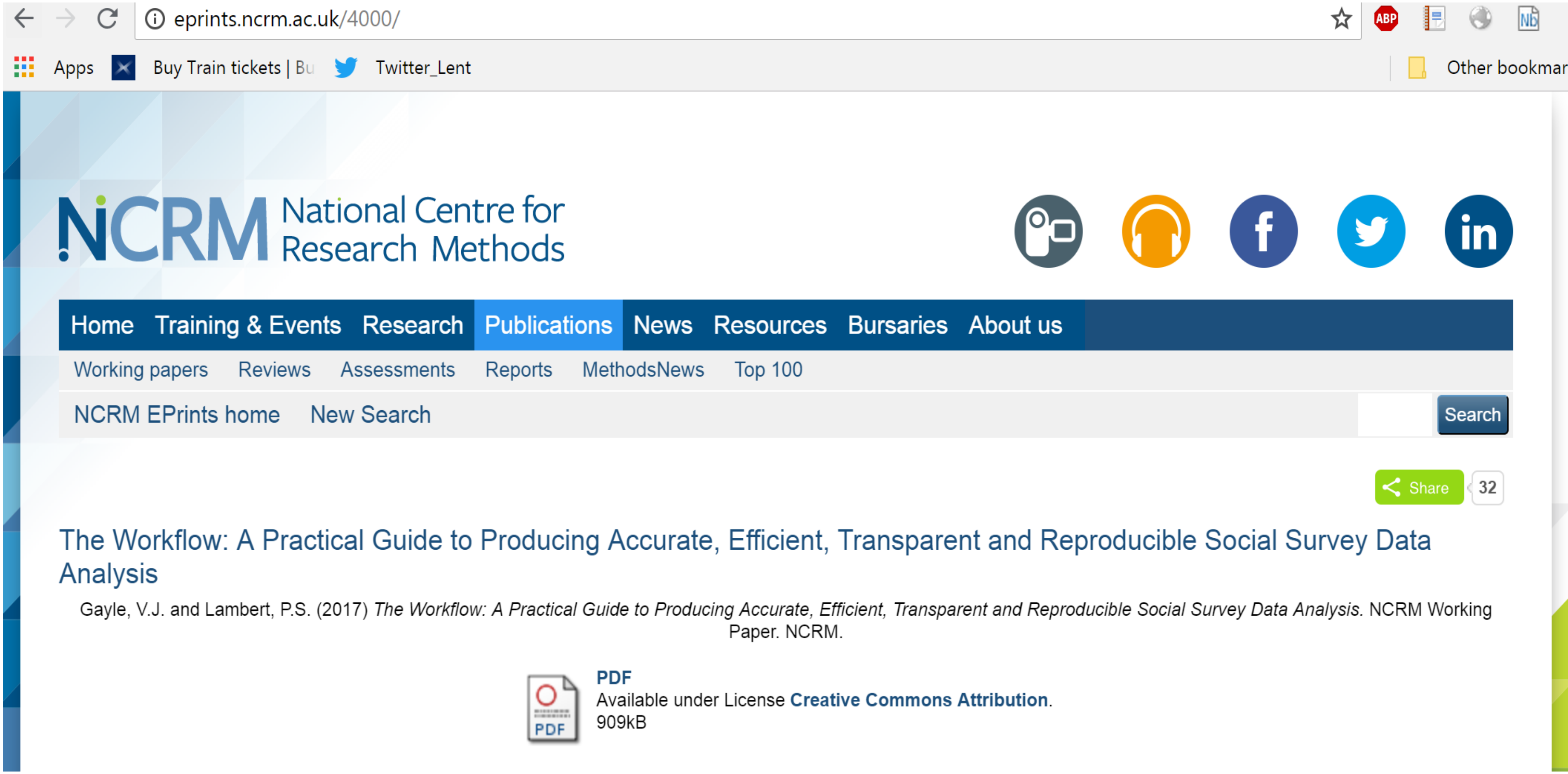
Therefore there is no choice other than to develop more sophisticated analyses which include more explanatory variables (i.e. statistical models)

Final Thought…

If applied research was easy, theorists would do it. But it is not as hard as the dense pages of Econometrica might lead you to believe. Avoid embarrassment by being your own best sceptic. And, especially, Don't Panic!

Angrist and Pischke (2008) *Mostly Harmless Econometrics*

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

THE MAIN EVENT!

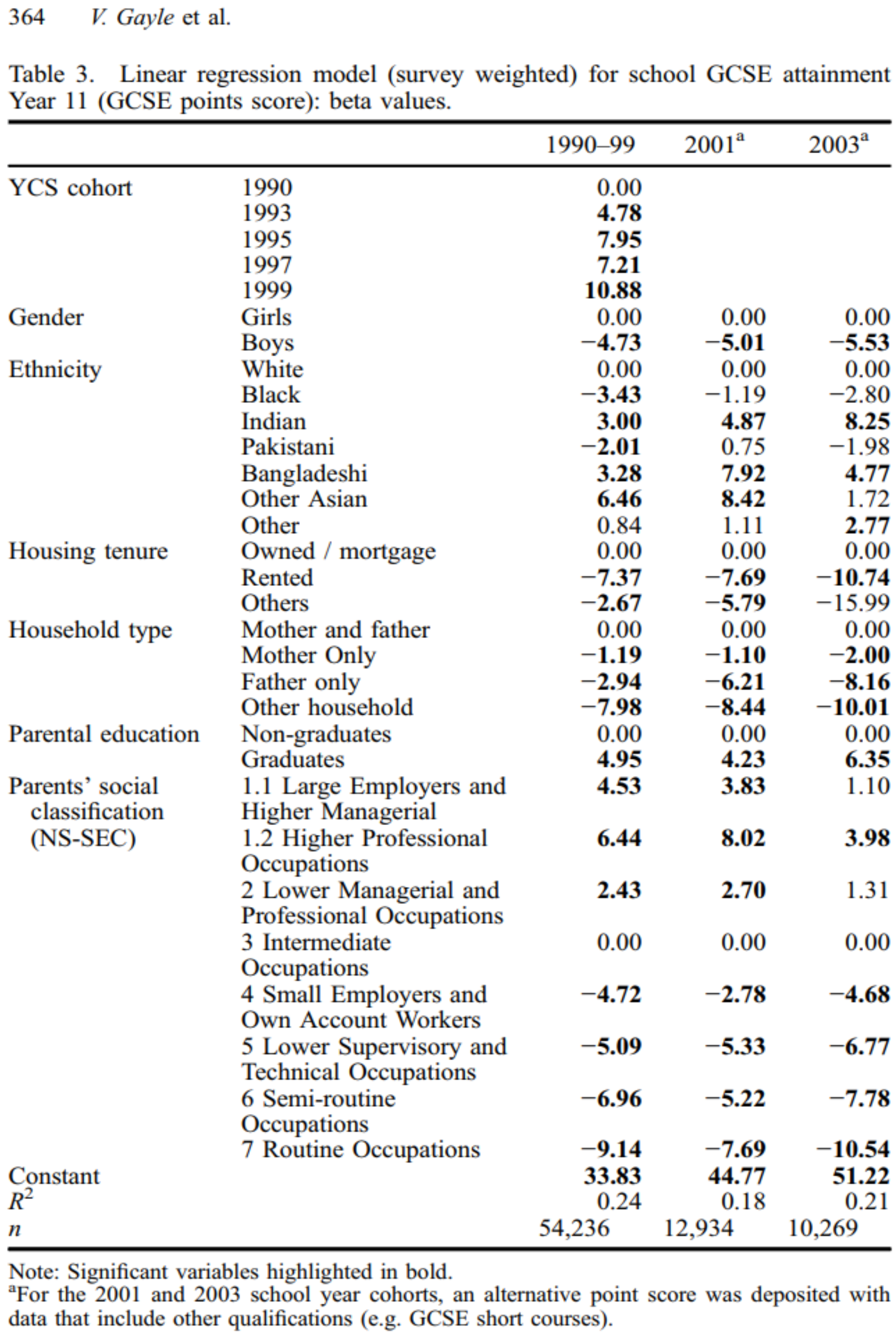


Table 3. Linear regression model (survey weighted) for school GCSE attainment Year 11 (GCSE points score): beta values.

		1990–99	2001[a]	2003[a]
YCS cohort	1990	0.00		
	1993	**4.78**		
	1995	**7.95**		
	1997	**7.21**		
	1999	**10.88**		
Gender	Girls	0.00	0.00	0.00
	Boys	**−4.73**	**−5.01**	**−5.53**
Ethnicity	White	0.00	0.00	0.00
	Black	**−3.43**	−1.19	−2.80
	Indian	**3.00**	**4.87**	**8.25**
	Pakistani	**−2.01**	0.75	−1.98
	Bangladeshi	**3.28**	**7.92**	**4.77**
	Other Asian	**6.46**	**8.42**	1.72
	Other	0.84	1.11	**2.77**
Housing tenure	Owned / mortgage	0.00	0.00	0.00
	Rented	**−7.37**	**−7.69**	**−10.74**
	Others	**−2.67**	**−5.79**	−15.99
Household type	Mother and father	0.00	0.00	0.00
	Mother Only	**−1.19**	**−1.10**	**−2.00**
	Father only	**−2.94**	**−6.21**	**−8.16**
	Other household	**−7.98**	**−8.44**	**−10.01**
Parental education	Non-graduates	0.00	0.00	0.00
	Graduates	**4.95**	**4.23**	**6.35**
Parents' social classification (NS-SEC)	1.1 Large Employers and Higher Managerial	**4.53**	**3.83**	1.10
	1.2 Higher Professional Occupations	**6.44**	**8.02**	**3.98**
	2 Lower Managerial and Professional Occupations	**2.43**	**2.70**	1.31
	3 Intermediate Occupations	0.00	0.00	0.00
	4 Small Employers and Own Account Workers	**−4.72**	**−2.78**	**−4.68**
	5 Lower Supervisory and Technical Occupations	**−5.09**	**−5.33**	**−6.77**
	6 Semi-routine Occupations	**−6.96**	**−5.22**	**−7.78**
	7 Routine Occupations	**−9.14**	**−7.69**	**−10.54**
Constant		**33.83**	**44.77**	**51.22**
R^2		0.24	0.18	0.21
n		54,236	12,934	10,269

Note: Significant variables highlighted in bold.
[a]For the 2001 and 2003 school year cohorts, an alternative point score was deposited with data that include other qualifications (e.g. GCSE short courses).

Table 3 p.364 in
Vernon Gayle, Susan Murray & Roxanne Connelly (2016) Young people and school General Certificate of Secondary Education attainment: looking for the 'missing middle', British Journal of Sociology of Education, 37:3, 350-370,
DOI: 0.1080/01425692.2014.935292

Part 15 Extra Material

We will only go over these slides if there is a consensus that the earlier material has been understood.

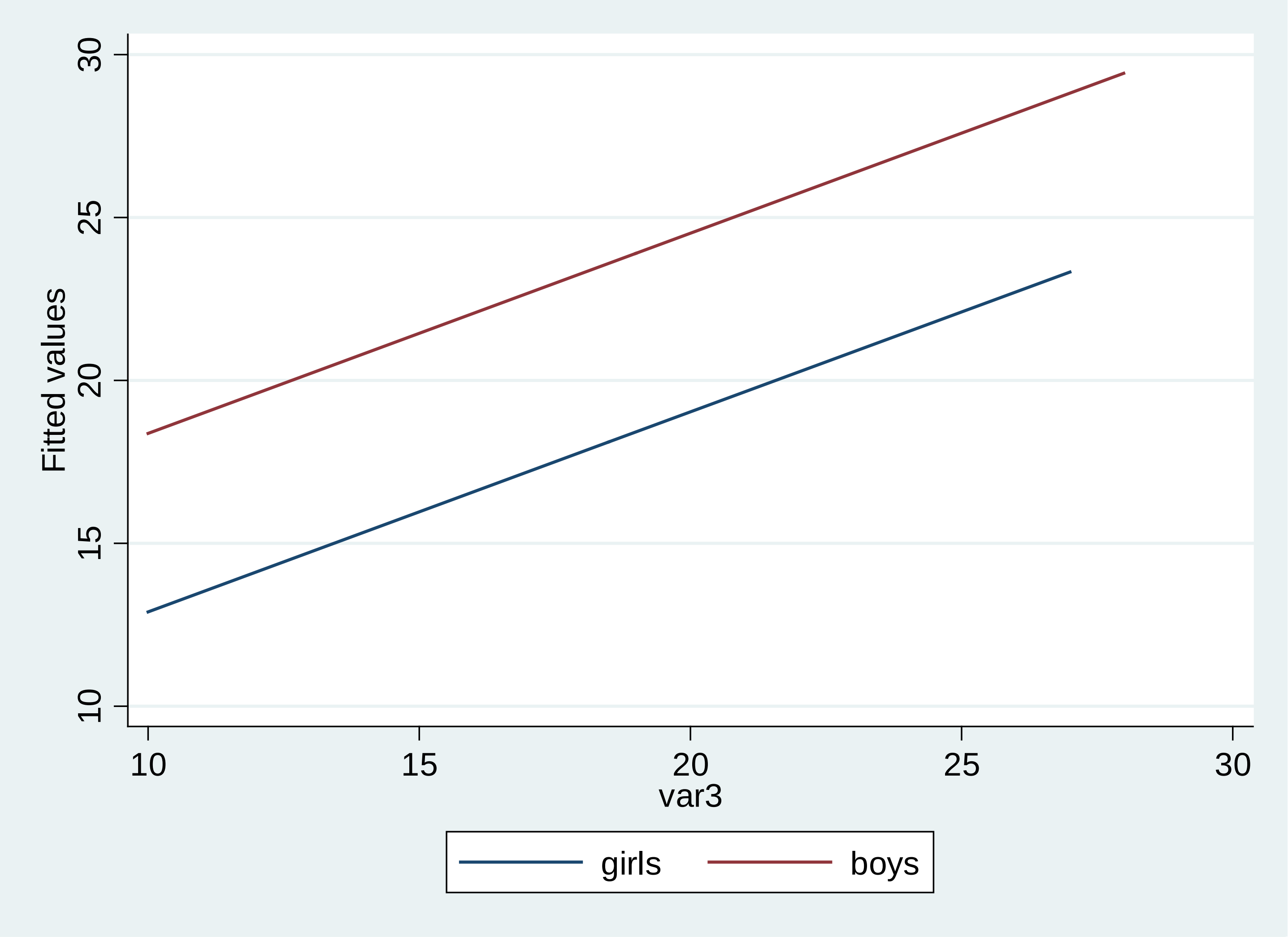
```
Survey: Linear regression

Number of strata   =         1                Number of obs     =     54,236
Number of PSUs     =    54,236                Population size   =  54,297.36
                                              Design df         =     54,235
                                              F(  24,  54212)   =     635.66
                                              Prob > F          =     0.0000
                                              R-squared         =     0.2362

------------------------------------------------------------------------------
             |             Linearized
    t0score2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cohort93 |   4.784237     .219699    21.78   0.000     4.353625    5.214849
    cohort95 |   7.948087    .2269095    35.03   0.000     7.503342    8.392831
    cohort97 |   7.208348     .227522    31.68   0.000     6.762403    7.654292
    cohort99 |   10.88168    .2375638    45.81   0.000     10.41605    11.34731
        boys |  -4.731582    .1482935   -31.91   0.000    -5.022238   -4.440926
       black |  -3.434908    .6481305    -5.30   0.000    -4.705248   -2.164567
      indian |   2.996469    .4929565     6.08   0.000      2.03027    3.962667
   pakistani |  -2.013062    .6400059    -3.15   0.002    -3.267479   -.7586456
 bangladeshi |   3.277848    1.279095     2.56   0.010     .7708115    5.784884
      oasian |    6.45998    .8873983     7.28   0.000     4.720673    8.199288
       other |   .8396727    .8057203     1.04   0.297    -.7395454    2.418891
      rented |  -7.368999    .2146037   -34.34   0.000    -7.789624   -6.948374
      ohouse |  -2.671727    .6382554    -4.19   0.000    -3.922712   -1.420741
     mumonly |  -1.190839    .2416689    -4.93   0.000    -1.664512    -.717166
     dadonly |  -2.935684    .4424884    -6.63   0.000    -3.802964   -2.068403
         ohh |  -7.976768     .537421   -14.84   0.000    -9.030117   -6.923419
     gradpar |   4.948721    .2194043    22.56   0.000     4.518687    5.378756
     nssec11 |   4.526175    .3648318    12.41   0.000     3.811102    5.241248
     nssec12 |   6.444145    .3303883    19.50   0.000     5.796581    7.091708
      nssec2 |   2.425388    .2433746     9.97   0.000     1.948372    2.902404
      nssec4 |  -4.715562    .2567644   -18.37   0.000    -5.218822   -4.212301
      nssec5 |  -5.087159    .3287678   -15.47   0.000    -5.731546   -4.442772
      nssec6 |  -6.963177    .2750893   -25.31   0.000    -7.502355      -6.424
      nssec7 |  -9.142231    .3210847   -28.47   0.000    -9.771559   -8.512902
       _cons |   33.82979    .2351548   143.86   0.000     33.36888    34.29069
------------------------------------------------------------------------------
```

Interaction Effects

‘While sometimes used in the broad sense of effects not operating separately, in statistical discussions it is typically restricted to effects that do not act additively on some response variable’ *Oxford Dictionary of Statistical Terms* p.203



The effect of variable var3 and the effect of gender – parallel lines

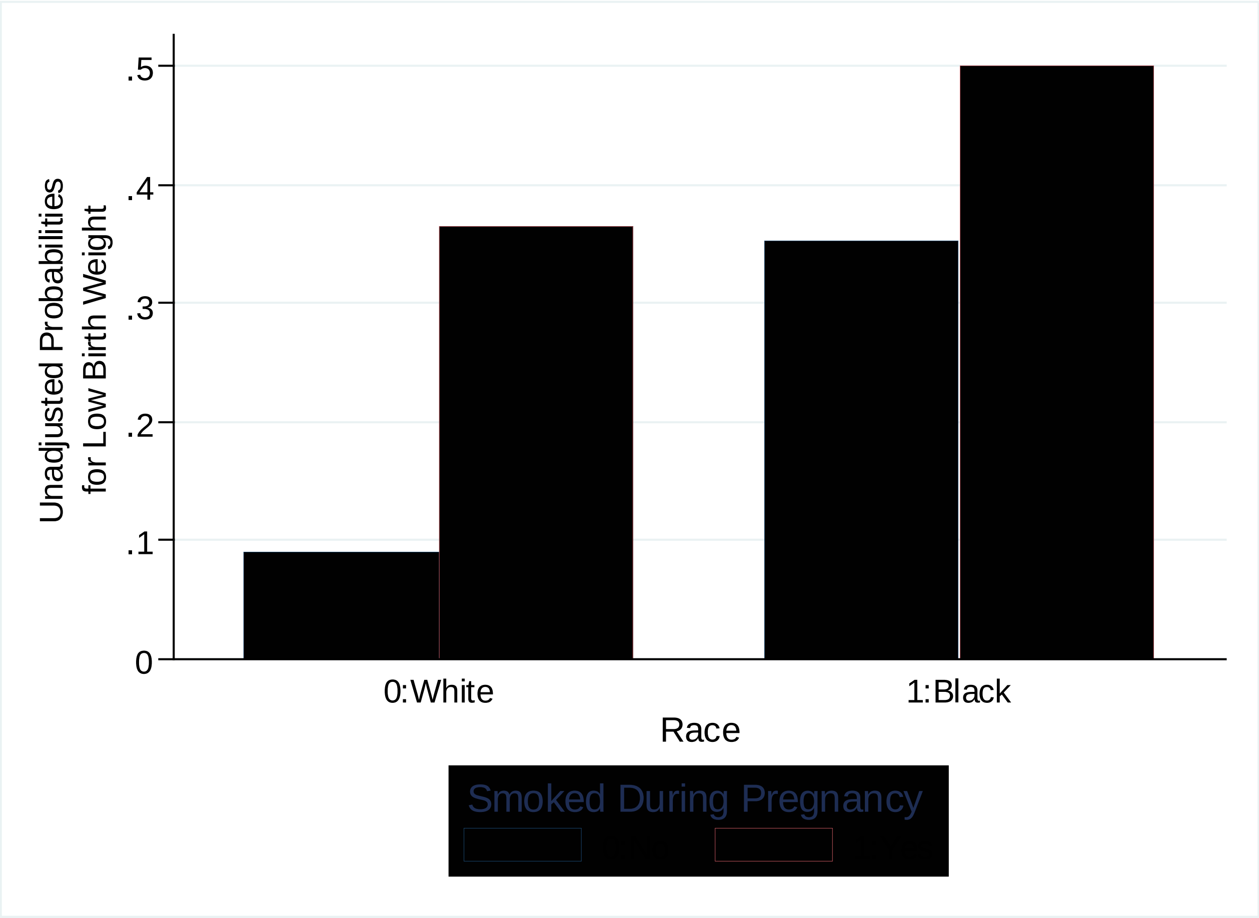
Interaction Effects

An interaction effect – when the effect of X_2 is also contingent on the value of X_1

Alternatively….

The effect of X_2 is not uniform across X_1

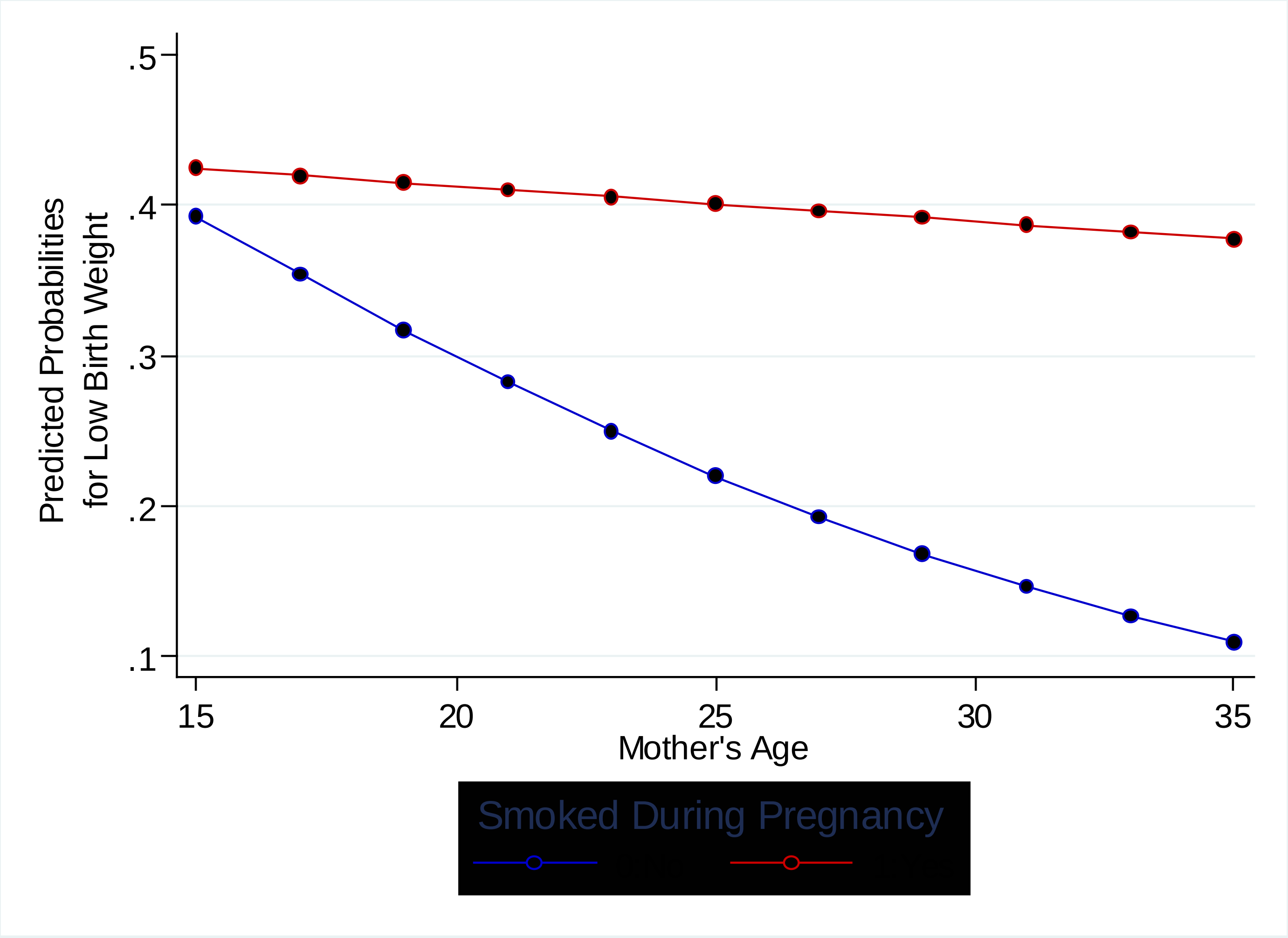
Low birth weight and interaction between smoking and race



p =.033

Neat example from Joanne Garrett, Robert Wood Johnson Clinical Scholars Program, University of North Carolina at Chapel Hill.

Low birth weight and interaction between smoking and age (years)



p=0.05

Neat example from Joanne Garrett, Robert Wood Johnson Clinical Scholars Program, University of North Carolina at Chapel Hill.

Professor Vernon Gayle

vernon.gayle@ed.ac.uk
@Profbigvern
github.com/vernongayle

AQMEN

THE UNIVERSITY
of EDINBURGH