Spring into Longitudinal Data Analysis

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh
vernon.gayle@ed.ac.uk
@profbigvern

April 2018

https://github.com/vernongayle/spring_into_longitudinal_data_analysis

(Re)-Introduction to Statistically Orientated Data Analysis

Part 1 Basic Statistical Concepts

Basic Concepts (probably revision)

- Variables - measures of social science concepts

- Cases - distinctive entities
  - People, firms, farms, hospitals, schools, local authorities, regions, nation states, animals, bee hives

Variables

- Outcome variables

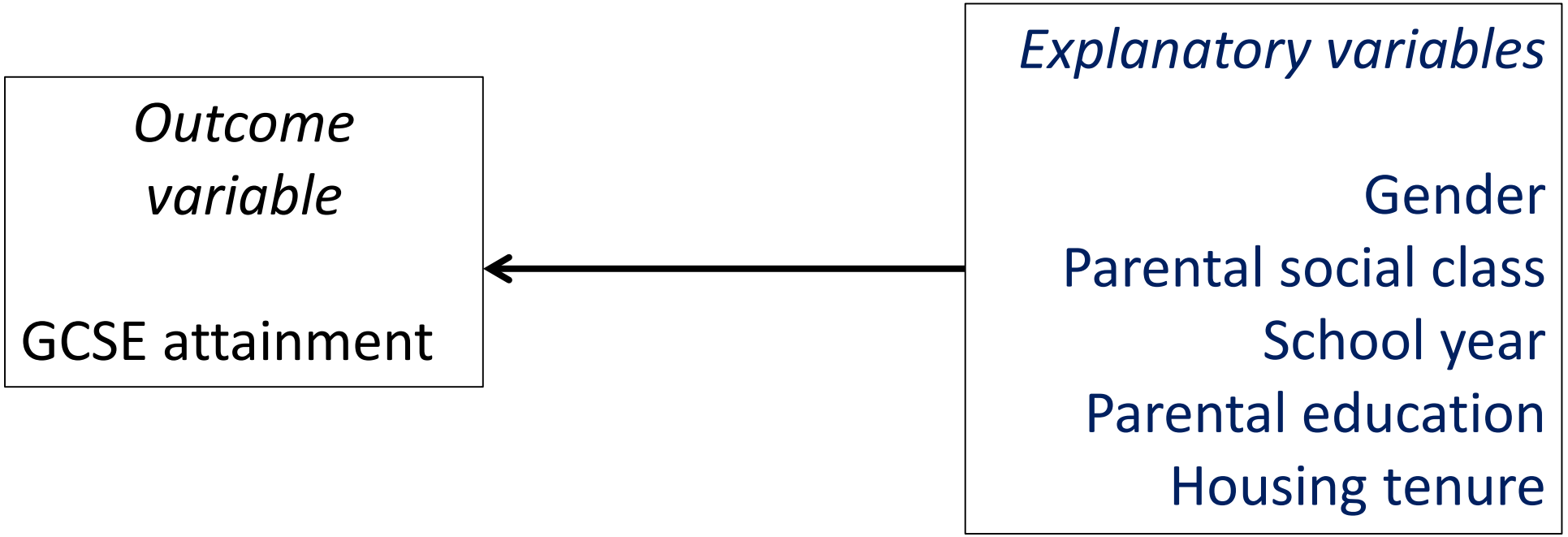
  (Y variables; dependent variables; dv)

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

- Explanatory variables (these variables explain outcome variables
  (X variables; X vars; Independent variables; IV)

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

Examples of variables in a real paper



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

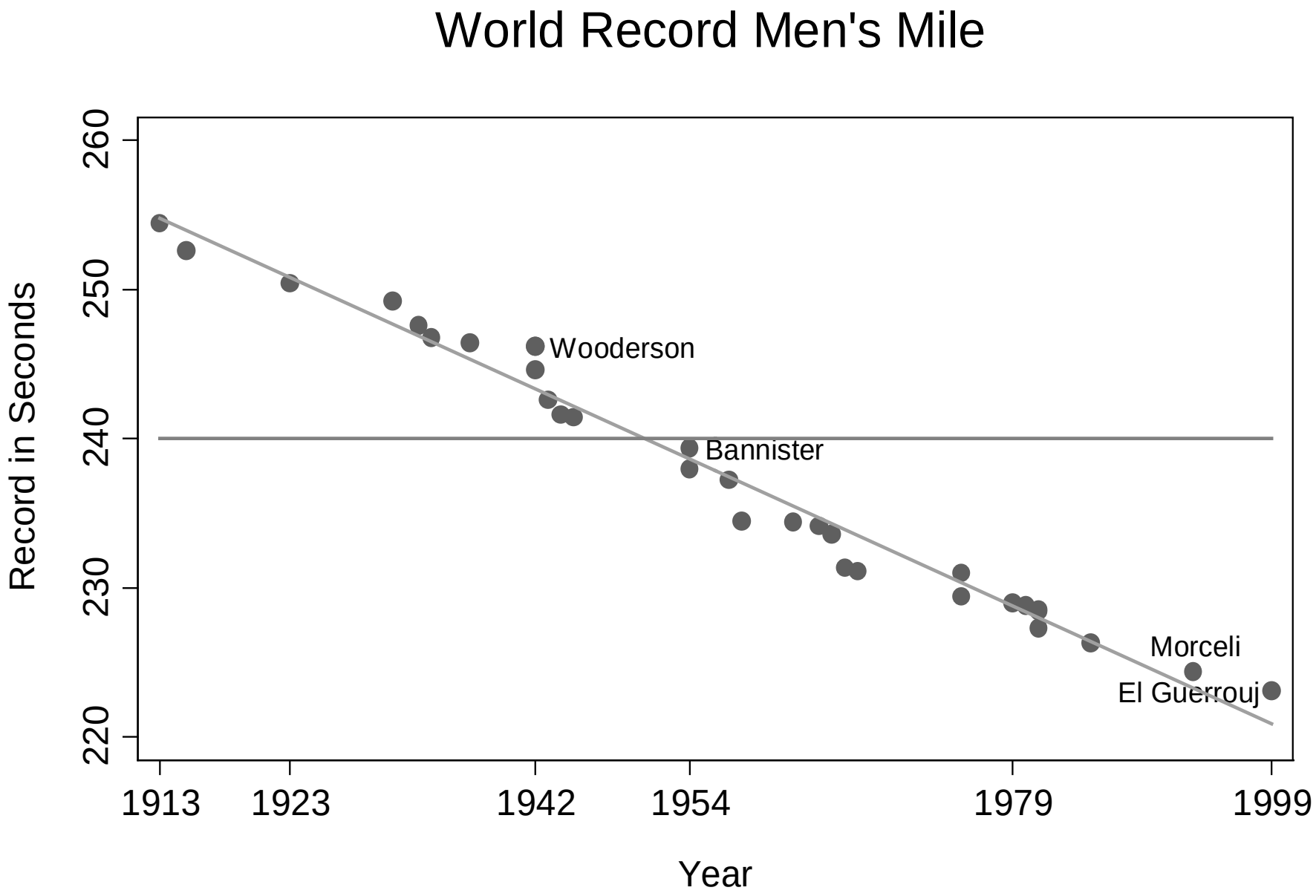
- **Univariate** – a single variable

- **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)

Part 2 Describing Data



Data Source: Wikipedia; Note: 4 minutes = 240 seconds

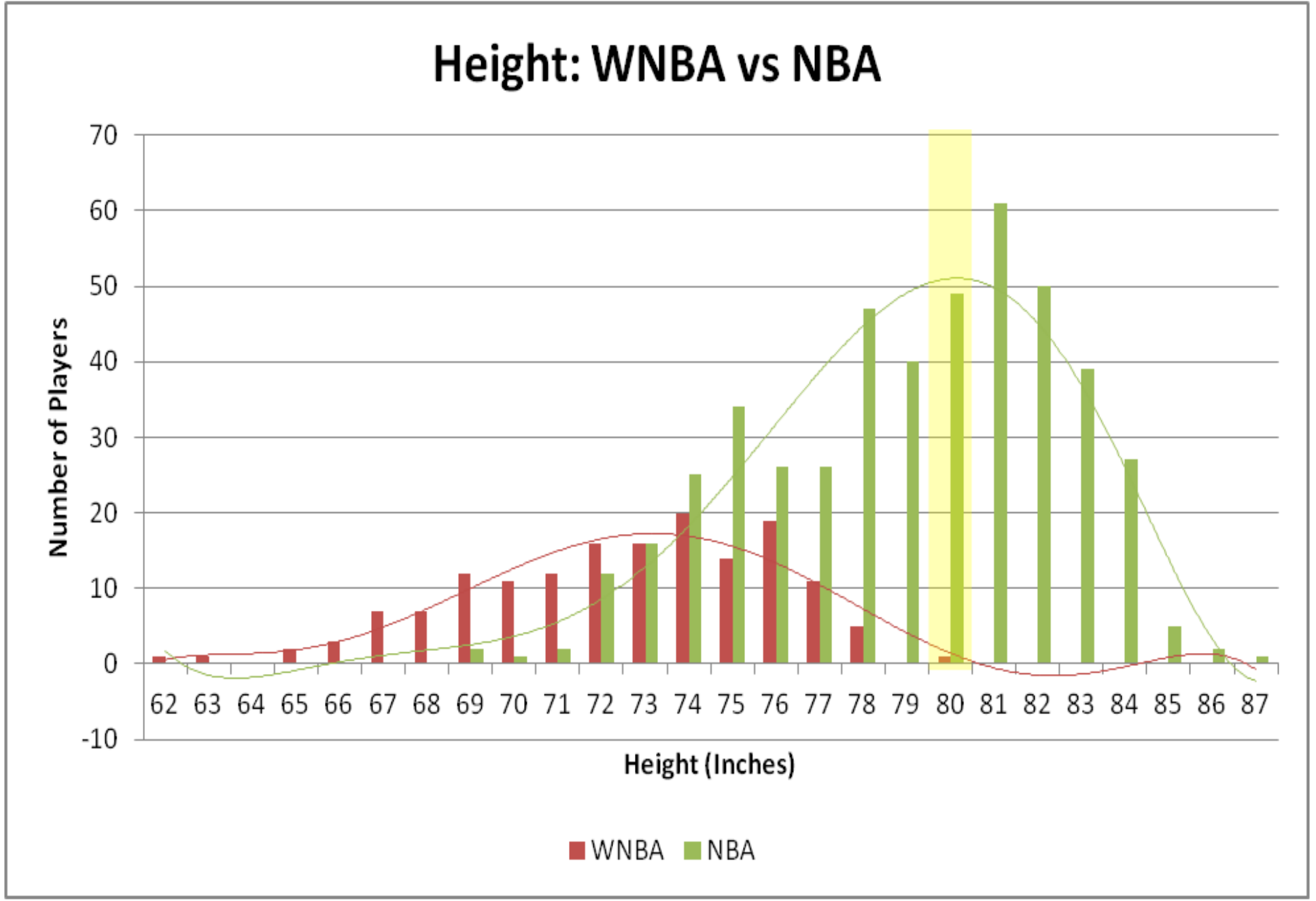
The Median

203 cm
193 cm
193 cm
191 cm
188 cm

---

185 cm **The 50th Percentile**

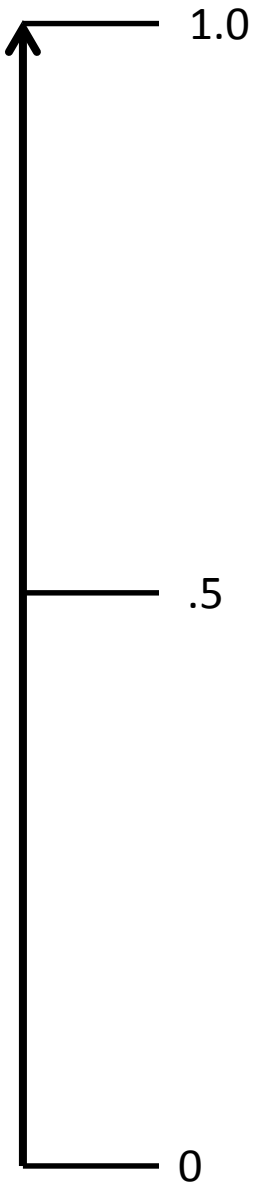
183 cm
180 cm
175 cm
173 cm
168 cm



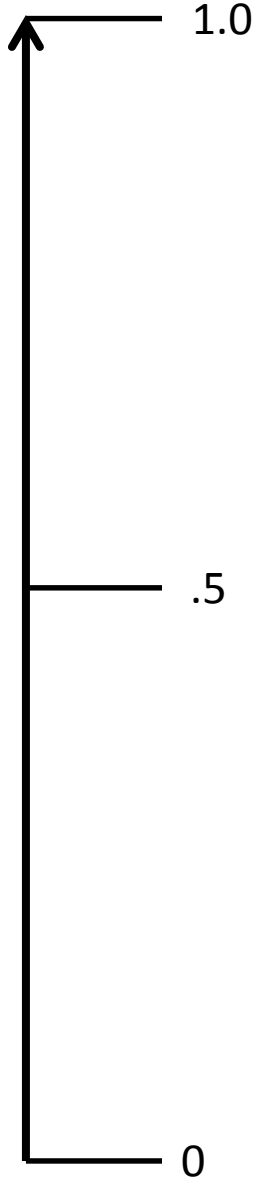
http://wagesofwins.com/2013/04/10/could-brittney-griner-make-it-in-the-nba/

Part 3 Probability

Probabilities take on values between 0 and 1

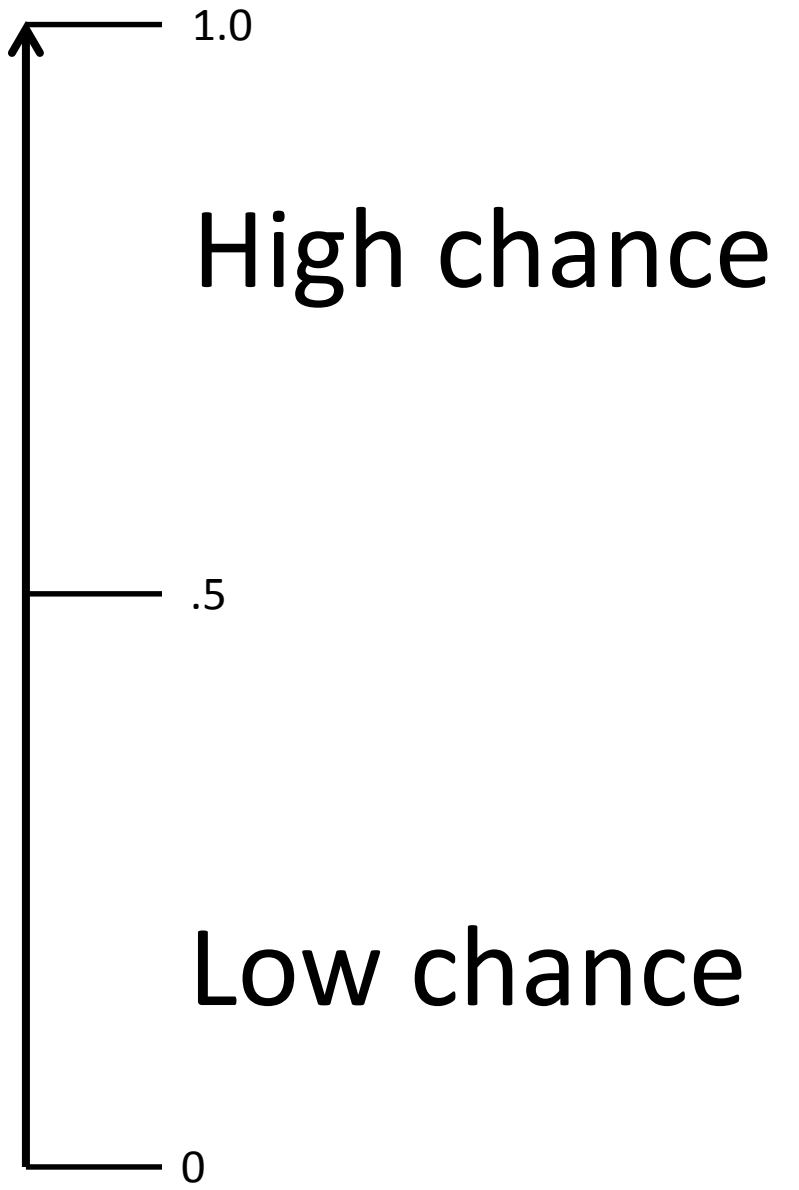


Denoted a “p”



Event will definitely occur

Event will definitely **not** occur


1.0
High chance
.5
Low chance
0

p=.50 even chance

Possible Outcomes

Total Outcomes

Possible Outcomes / Total Outcomes

1 Ace of Spades / 52 Cards in Total

In Pairs Guess the Probability?

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?

When I toss my coin 5 times what is the chance of 5 heads in a row?

½ x ½ x ½ x ½ x ½ = ??

Think ½ x ½ = ¼ (i.e. half of a half)

			Toss 4	p (all_heads)
0.5				0.5
0.5	0.5			0.25
0.5	0.5	0.5		0.125
0.5	0.5	0.5	0.5	0.0625

				Toss 5	p (all_heads)
0.5					0.5
0.5	0.5				0.25
0.5	0.5	0.5			0.125
0.5	0.5	0.5	0.5		0.0625
0.5	0.5	0.5	0.5	0.5	0.03125

Part 4 *p* Values

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			

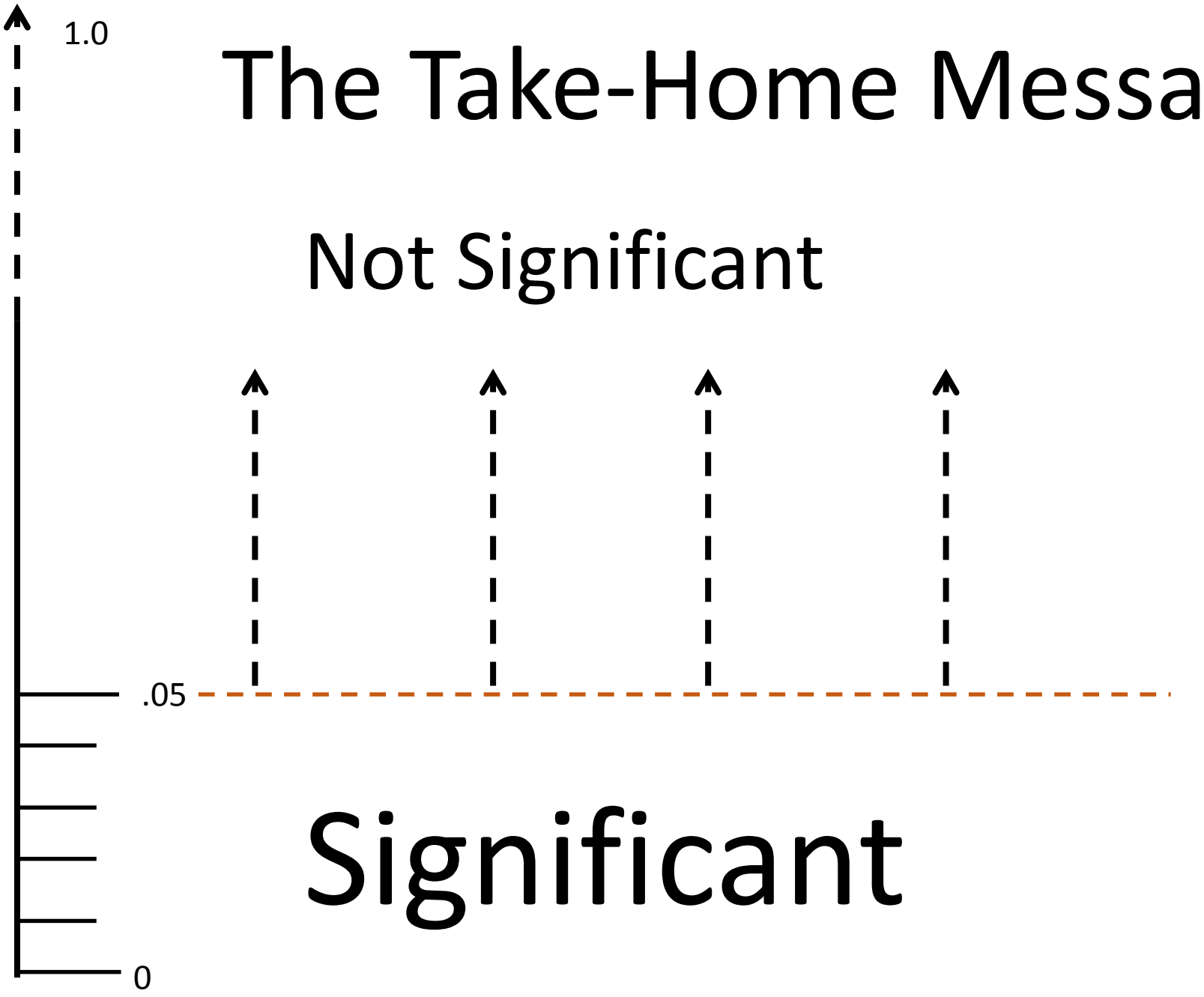
p values

	(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

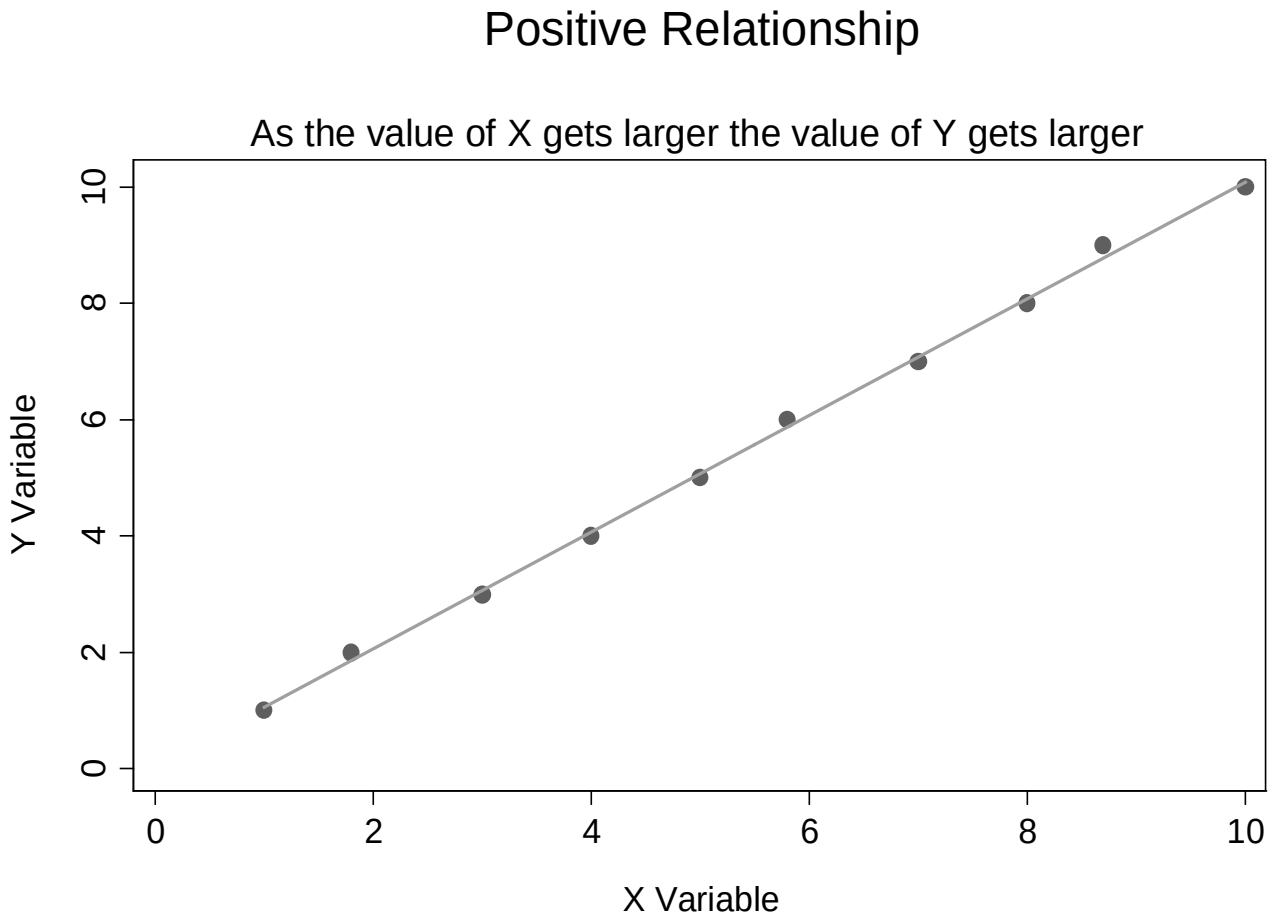
1 in a million - 50p tossed 20 times all coming out heads

Part 5 Significance Tests & Probability

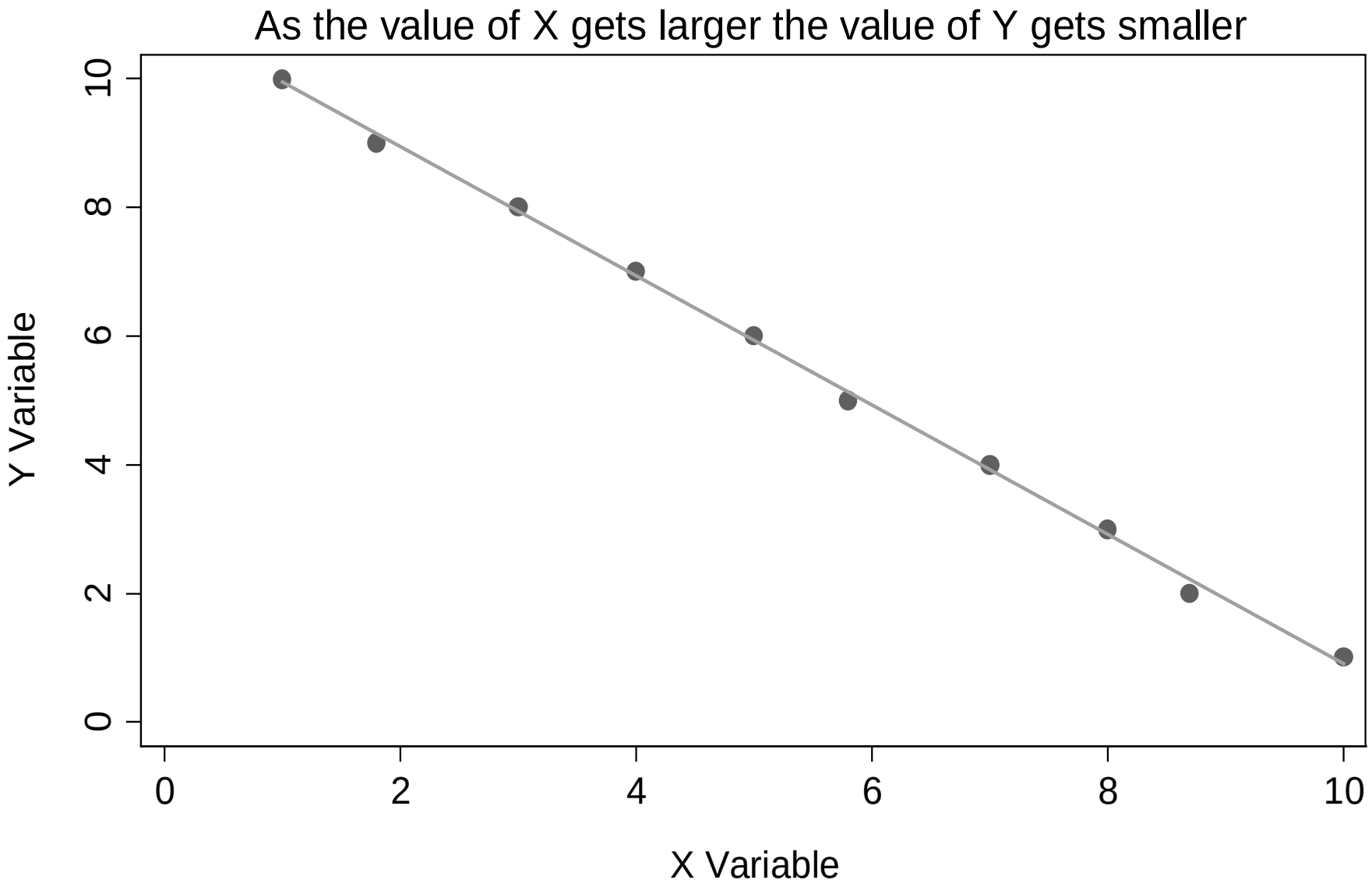
The Take-Home Message



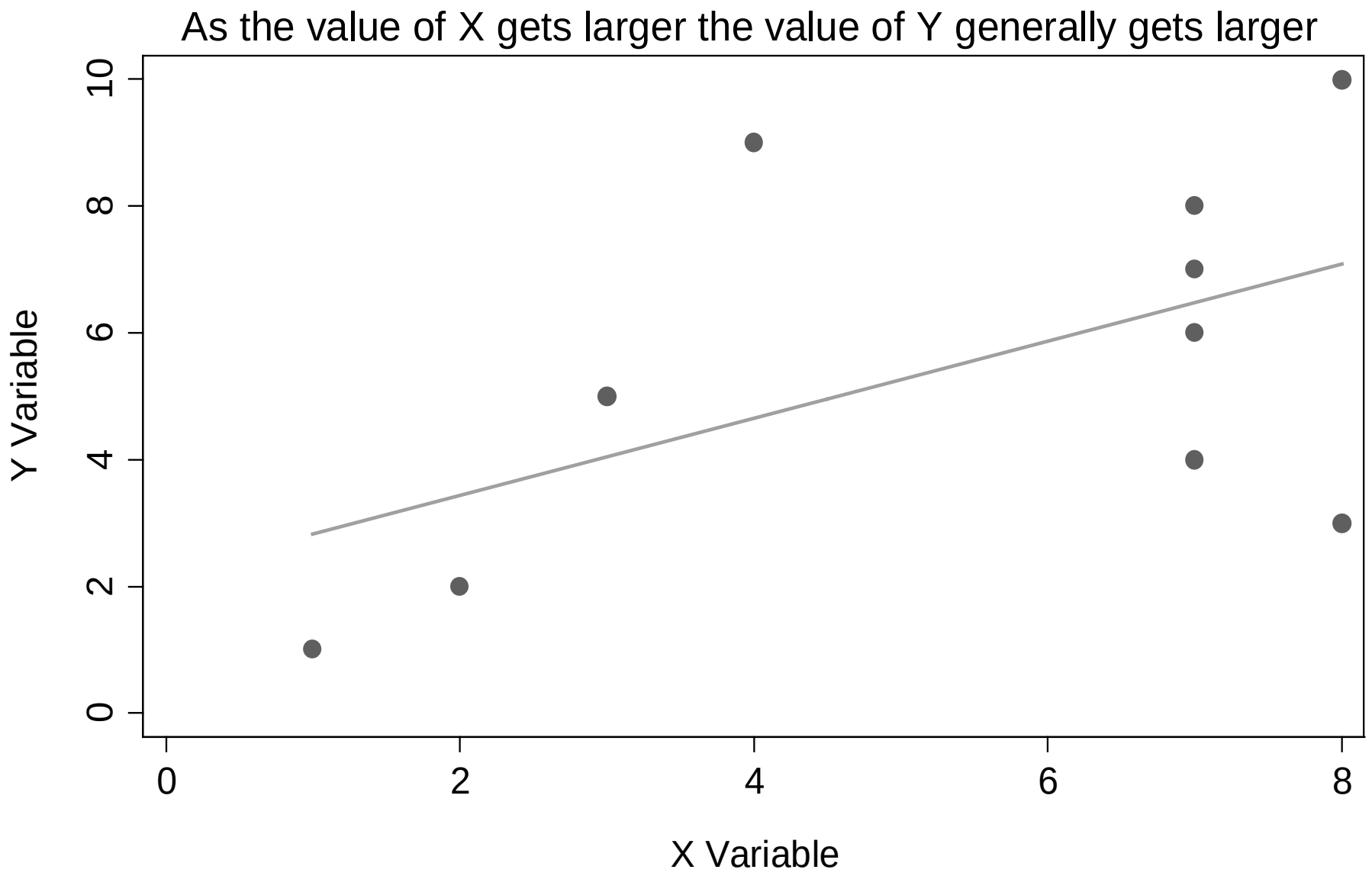
Part 6 Relationships in Data (probably revision)



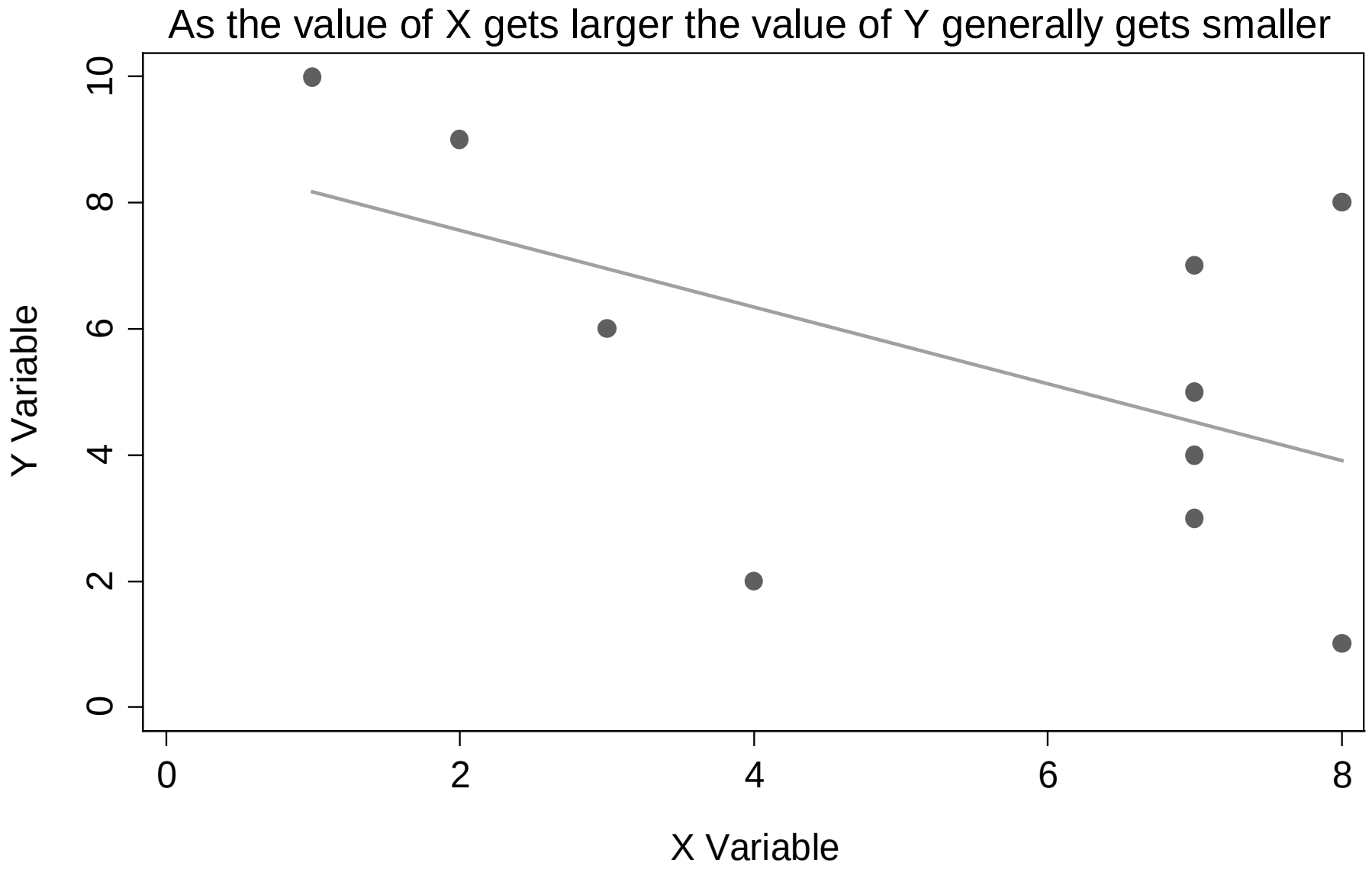
Negative Relationship



Weaker Positive Relationship

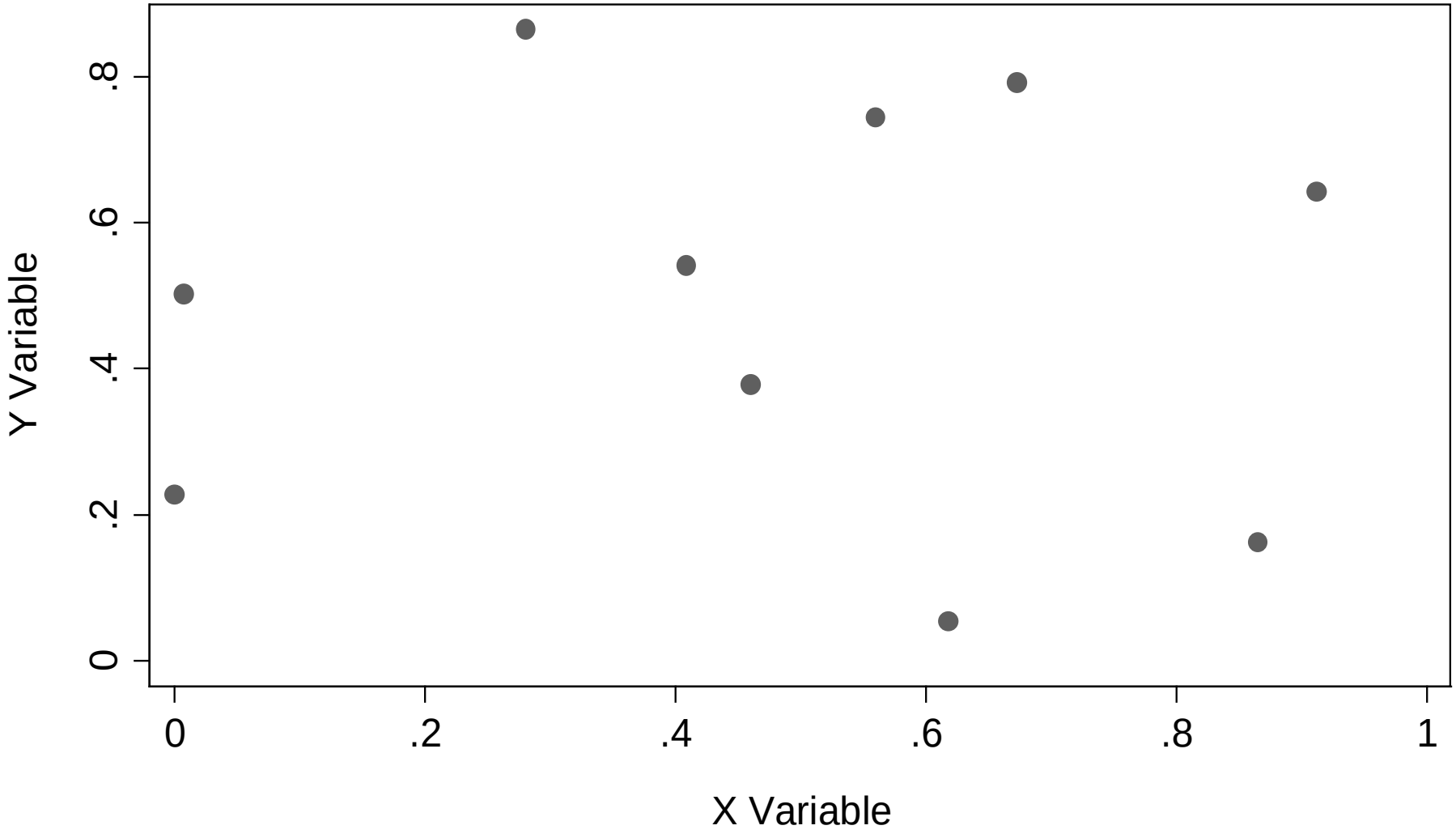


Weaker Negative Relationship

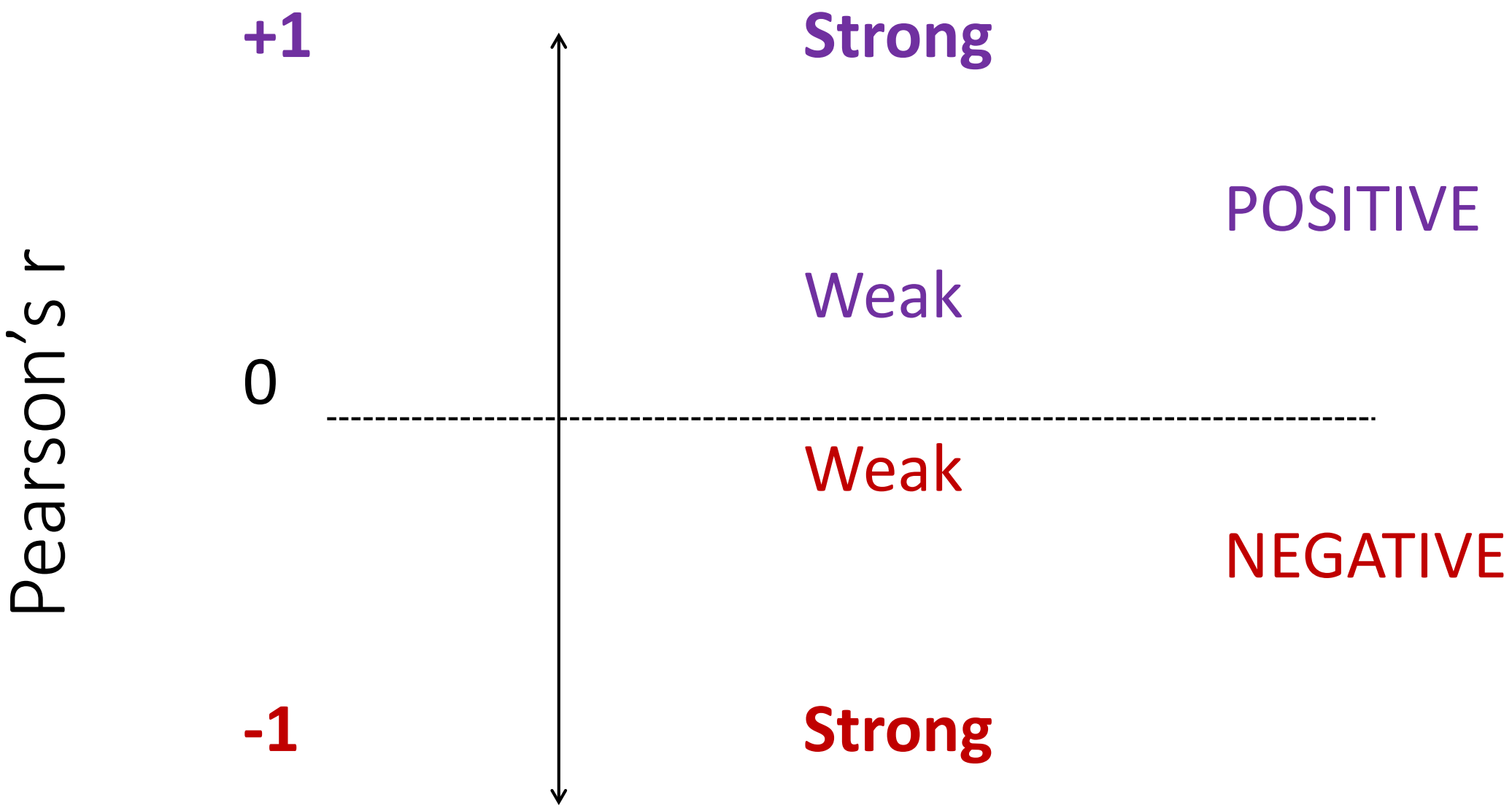


No (linear) Relationship

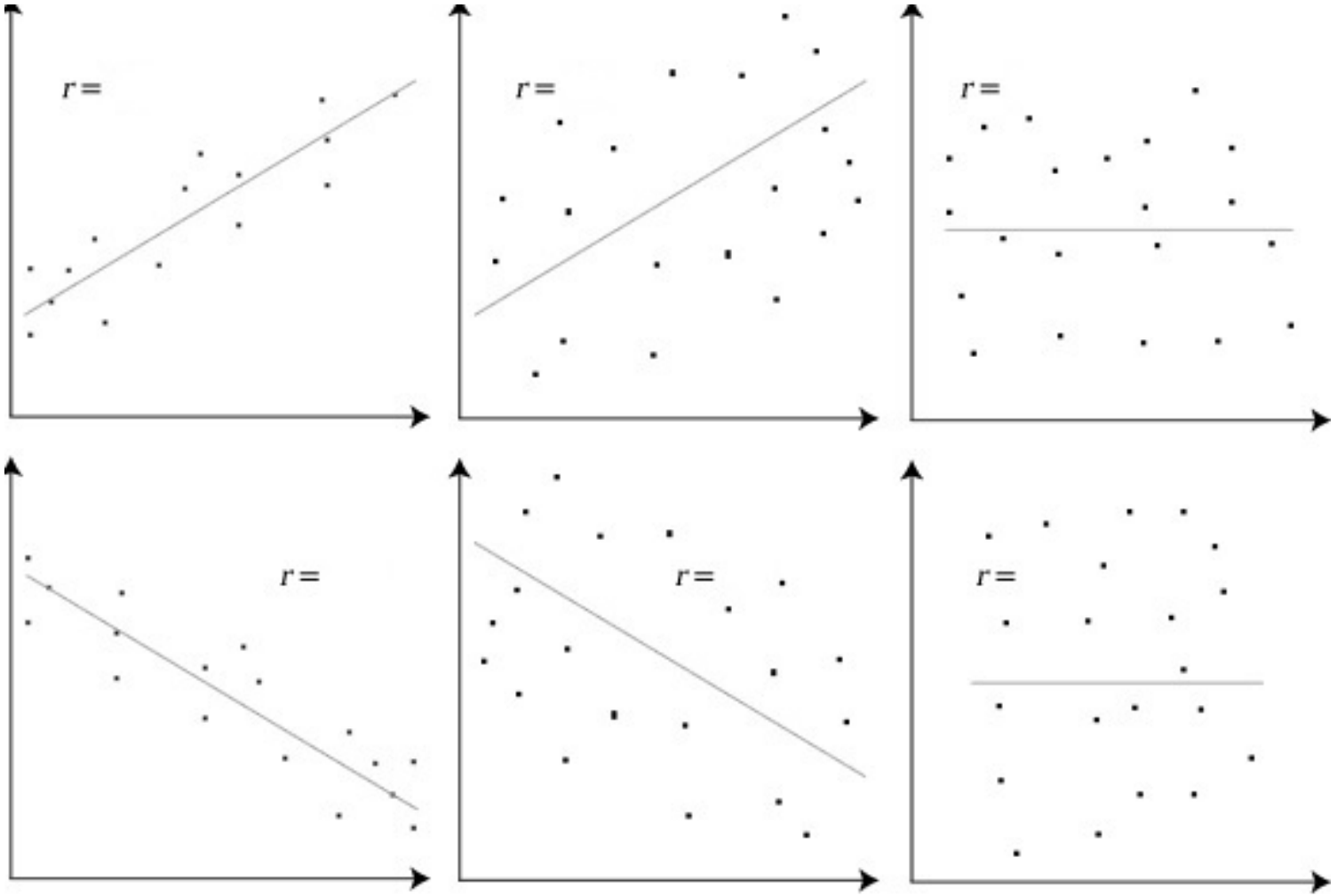
As the value of X gets larger
any guess as to what happens to the value of Y



Part 7 Correlations



What is the value of *r* ?

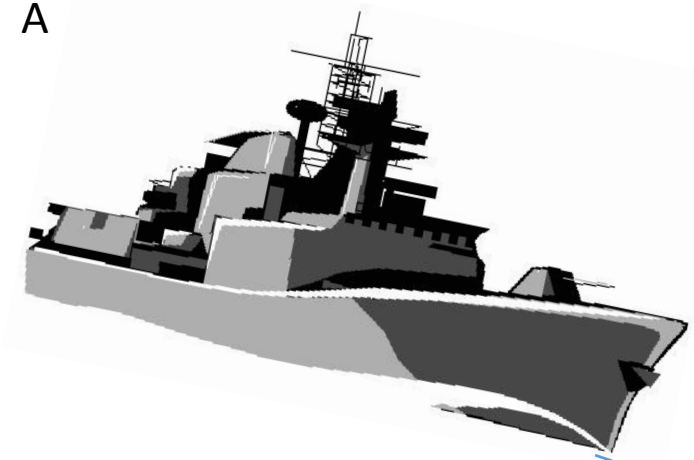


Part 8 Confidence Intervals

95% Confidence intervals around mean :

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

A

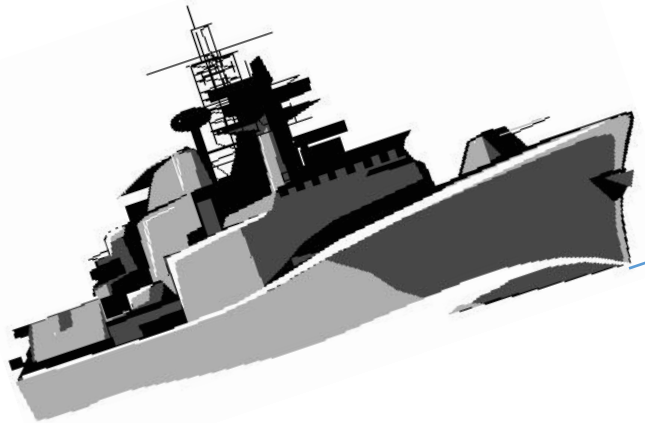


Ship A plans to be at point C at 10:00 am

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

C

B

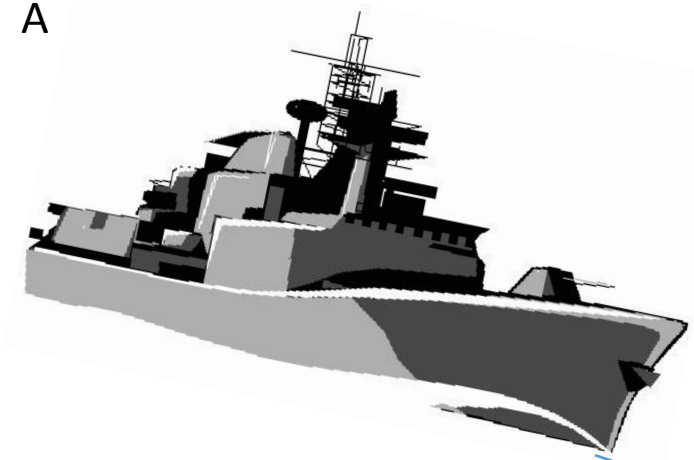


Ship B plans to be at point C at 10:15 am

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

On Another Day…

A

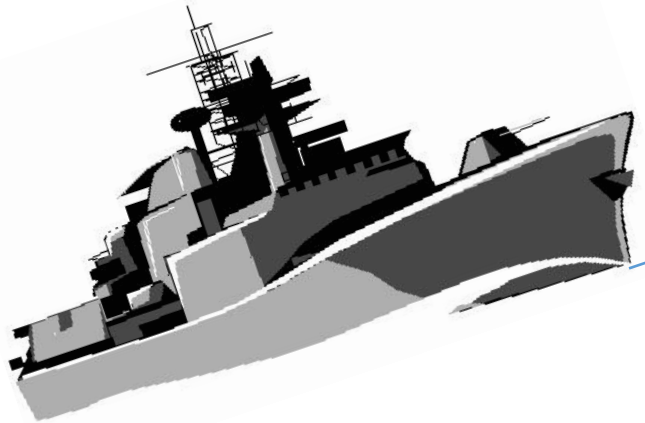


Ship A plans to be at point C at 10:00 am

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

C

B



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

DAY 2

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh
vernon.gayle@ed.ac.uk
@profbigvern

April 2018

Introduction to Longitudinal Data

Why Use Longitudinal Data?

- UK has an unparalleled collection
- These resources are critical for analysing social change (and social stability)
- But they need justification because they are costly in money and time

Longitudinal Social Surveys

- Cross-sectional data
  - Respondents surveyed at only one time point
- Longitudinal data
  - Repeated contacts (with the same individuals)
  - Respondents surveyed at multiple time points

The Up Series

https://en.wikipedia.org/wiki/Up_Series

https://www.youtube.com/watch?v=VVQ96wfbf_0

Study Designs

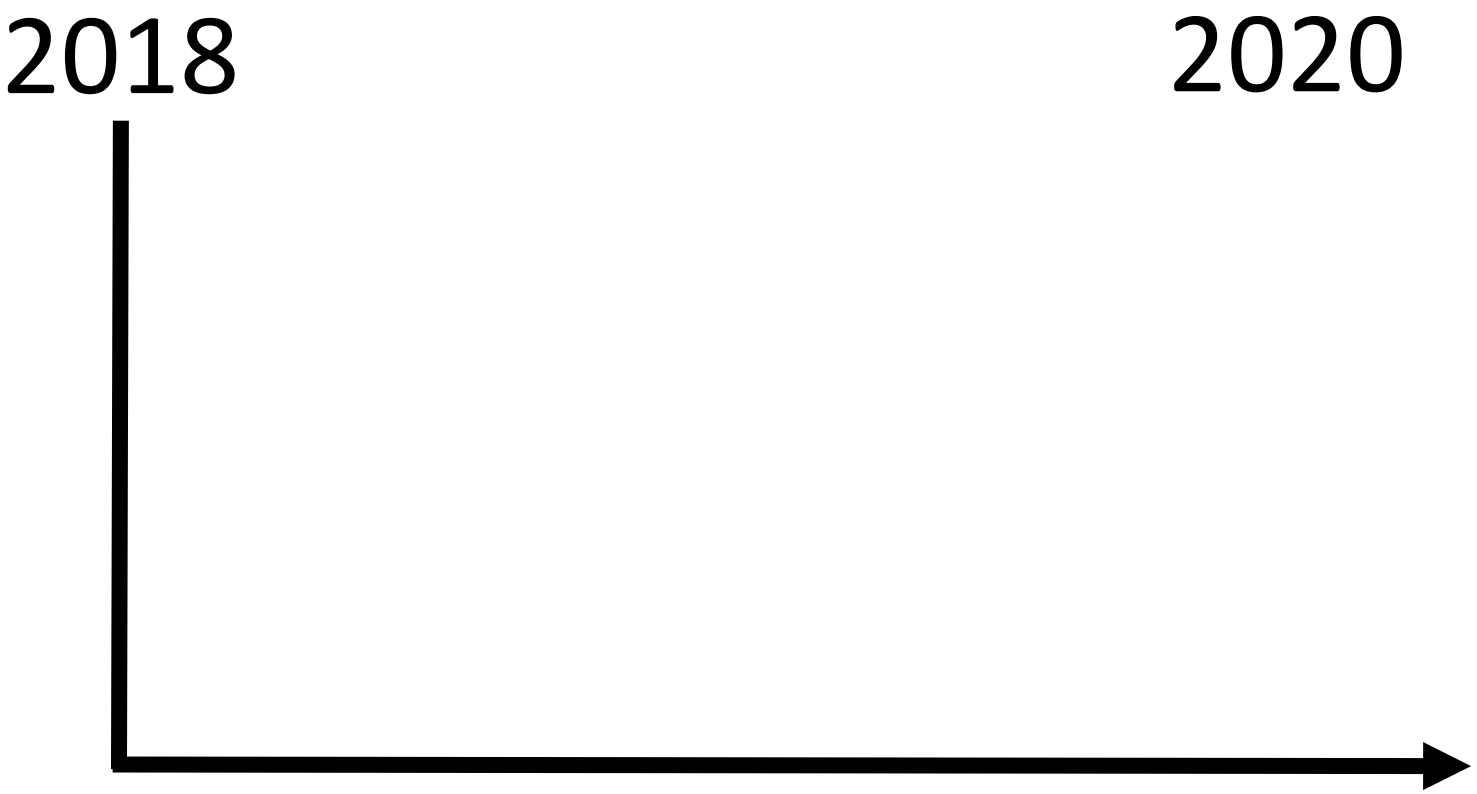
Retrospective Measures

Town where born?

Father’s age when you were born?

Mother worked full-time when started secondary school?

Study Designs



Prospective Designs (contemporaneous measures)

Did paid work last week

Current (legal) marital status

Belongs to a religion

Commutes to work on a skateboard

Longitudinal Social Science Study Designs

Panel Study

The panel are the group and are repeatedly studied

- US (PSID)
- Germany (SOEP)
- Britain BHPS/UKHLS
- Australia (HILDA)
- Canada (SLID)
- Swiss (SHP); Korea (KLIPS); Russia (RLMS)

Longitudinal Social Science Study Designs

Cohort Study

- Repeated contacts data collection

(simply a specific form of panel design in my view)

- Principally concerned with charting the development of a particular ‘group’ from a certain point in time

Longitudinal Social Science Study Designs

- Cohort Study
  - A birth cohort of babies born in a particular year (e.g. 1946; 1958; 1970; 2000-2)
  - A youth cohort, a group of pupils who completed compulsory education in the same year (YCS; LSYPE)

Research Using Longitudinal Social Survey Datasets

- For many social research projects cross-sectional data will be sufficient
- Most social research projects can be improved by the analysis of longitudinal data
- Some research questions require longitudinal data

Questions that Require Longitudinal Data

- Flows into and out of poverty
- The effects of family migration on the woman's subsequent employment activities
- Numerous policy intervention examples
- Numerous examples relating to 'individual' development

Key Messages (so far….)

- For many social research projects cross-sectional data will be sufficient
- Most social research projects can be improved by the analysis of longitudinal data
- *Researchers are likely to make more rapid progress using existing large-scale longitudinal data resources*

Key Messages (so far….)

- Some research questions require longitudinal data
- Longitudinal data are not a panacea

'This longitudinal study suggests that notwithstanding the dominant effect of severity of intellectual impairment, a number of factors within and outside the family may also contribute to higher attainment in reading, writing and numeracy.

In particular mainstream schooling for those with less severe disabilities appears to have benefited the children in this study' (p.390).

Turner, S., Alborz, A. and **Gayle, V.** (2008) 'Predictors of academic attainments of young people with Down's syndrome', *Journal of Intellectual Disability Research*, 52(5), pp. 380-392.

Subjective Well-Being & Happiness

- Non-economic measures of social progress
- “Improving the quality of our lives should be the ultimate target of public policies” Angel Gurría, OECD Secretary-General
- UK commitment to developing wider measures of well-being
- Tailoring government policies to the things that matter

- Moving house itself causes a boost in happiness, and brings people back to their initial levels
- Moving and set-point theory
- Long-distance migrants are at least as happy as short-distance migrants despite the higher social and psychological costs involved
- Re-theorize moving within a conceptual framework that accounts for social well-being from a life-course perspective

Nowok, B., van Ham, M., Findlay, A. and **Gayle, V.** (2013) 'Does migration make you happy? A longitudinal study of internal migration and subjective wellbeing', *Environment and Planning A*, 45(4), pp. 986-1002.

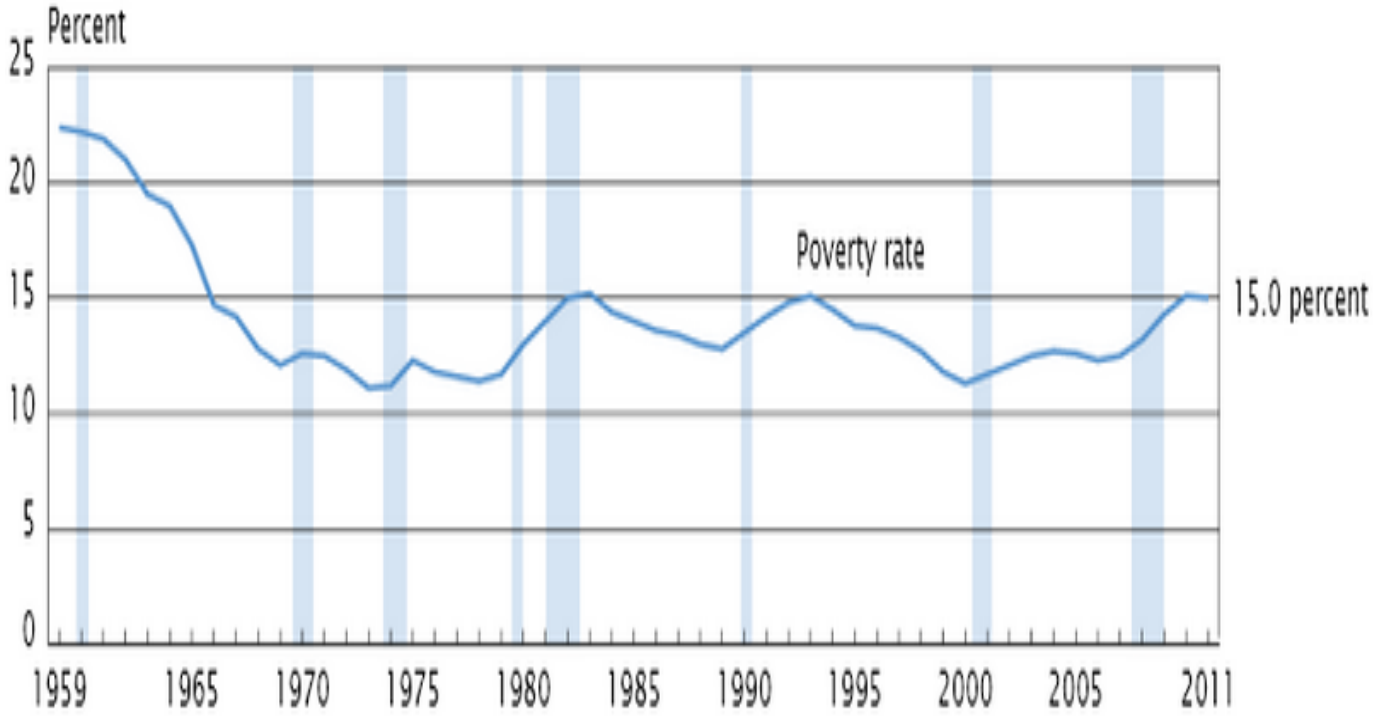
The Bigger Picture

- UKHLS is the largest living observatory of contemporary social life
- Contribution to the ‘evidence base’
- Contribution to empirically informed planning
- *Influencing behaviour and informing interventions*
- *Contributing to a fair and vibrant society*

Examples…

- Cohort studies - secondary smoking effects on children (back in the news)
- Whitehall Studies - influenced successive governments' thinking on social gradients in health
- Whitehall Studies -dispelled the myth that high status jobs have higher risk of heart disease

USA Poverty Rate 1959 - 2011



Note: The data points are placed at the midpoints of the respective years. For information on recessions, see Appendix A.

Source: U.S. Census Bureau, Current Population Survey, 1960 to 2012 Annual Social and Economic Supplements.

BHPS British Household Panel Survey

- Poverty rates flattened out in 1990s
- BHPS showed apparent cross-sectional stability but a hidden longitudinal flux
  - Substantial turnover or churning
  - The poor were not always poor
- Not detectable without panel data!

BHPS British Household Panel Survey

- UK Poverty rate approximately 18%
- In a 6 year periods one-third of individuals were poor at least once
- Only 2% were poor for all six years!
- Repeated short spells of poverty were more common than one long spell

The Consequences…

- Contributed to the ‘rubber band theory’
  - we are attached to an elastic tether
- Influenced the Labour government’s welfare reforms in the late 1990s
  - focussing on moving people into work and making work pay
- Now influences how living standards are measured in Britain
  - Official Statistics now include household panel based information

Summary Messages

- For many social research projects cross-sectional data will be sufficient
- Most social research projects can be improved by the analysis of longitudinal data
- Some research questions require longitudinal data

The Research Value of Longitudinal Data

Questions that Require Longitudinal Data

- Flows into and out of poverty
- The effects of family migration on the woman's subsequent employment activities
- Numerous policy intervention examples
- Numerous examples relating to 'individual' development

Methodological Benefits of Longitudinal Social Science Data

- Micro-level social processes
- Temporal ordering of events
- Improving control for residual heterogeneity
- Improving control for state dependence

Micro-Level Social Processes

- Cross-sectional data = a snap shot
  - Good for studying the immediate
  - Several datasets can study macro / or gross changes

- Repeated contacts data allow the study of
  - The passage of time
  - Individual (or household) change/stability
  - Processes that occur at the micro-level of the individual (or family)
  - Surprises (or shocks)

Temporal Ordering of Events (Direction of Influence)

- Time moves in one direction so…
  - An event in 1990 comes before an event in 1995
  - Experiences at primary school could affect university entry
  - Teenage smoking could influence health in old age

- But not *vice versa*
  - ➢ One sociology professor has argued with me suggesting that time does not move in only one direction

Temporal Ordering of Events (Direction of Influence)

- There is unequivocal evidence from cross-sectional data that, overall, the unemployed have poorer health
- This is consistent with both
  A. Unemployment causing ill health
  B. Ill health causing unemployment
- These two substantive stories are quite different

Month	Level of Health (20 = Good Health)	Employ Status
1	17	Employed
2	17	Employed
3	17	Employed
4	17	Unemployed
5	17	Unemployed
6	10	Unemployed
7	16	Unemployed
8	5	Unemployed
9	4	Unemployed
10	3	Unemployed
11	2	Unemployed
12	1	Unemployed

Person A

Became unemployed this has affected his level of health

Month	Level of Health (20 = Good Health)	Employ Status
1	17	Employed
2	1	Employed
3	1	Employed
4	1	Unemployed
5	1	Unemployed
6	1	Unemployed
7	1	Unemployed
8	1	Unemployed
9	1	Unemployed
10	1	Unemployed
11	1	Unemployed
12	1	Unemployed

Person B

Poor health led to unemployment

(because of poor job performance)

In a cross-sectional study (at month 12)

- Person A would have been unemployed for 9 months and have a health score of 1
- Person B would have been unemployed for 9 months and have a health score of 1
- This is an obvious example of how panel (i.e. repeated contacts) data can make an essential contribution to untangling social relationships

Improving Control for Omitted Explanatory Variables

- Residual Heterogeneity
  - Omitted explanatory variables
  - Unobserved heterogeneity
- The possibility of substantial variation between similar individuals due to unmeasured, and possibly immeasurable, variables is known as '*residual heterogeneity*'

Improving Control for Omitted Explanatory Variables

Because data collection instruments often fail to capture the detailed nature of social life there is, almost inevitably, considerable heterogeneity in response variables even amongst respondents that share the same characteristics across all of the explanatory variables

Improving Control for Omitted Explanatory Variables

As long as we make the assumption that (at least some of) these effects are enduring there are techniques for accounting for omitted explanatory variables if we have data at more than one time point

Improving Control for Omitted Explanatory Variables

- There are no routine methods of accounting for omitted explanatory variables in cross-sectional analysis
- It is sometimes claimed that the main advantage of longitudinal data is that it facilitates improved control for the plethora of variables that are omitted from any analysis
- Panel data won't completely sweep this problem away, but suitable models can improve control for, and estimate the effects of, residual heterogeneity

Improving Control for the Effects of Previous States (state dependence)

A frequently noted empirical regularity in the analysis of unemployment data is that those who were unemployed in the past or have worked in the past are more likely to be unemployed (or working) in the future

(Nobel Prize winner J.J. Heckman)

Improving Control for the Effects of Previous States (state dependence)

- Much of human behaviour is influenced by previous behaviour and outcomes (positive feedback)

- McGinnis (1968) '*axiom of cumulative inertia*'

Improving Control for the Effects of Previous States (state dependence)

- Working in May = more likely to be working in June
- Married this year = more likely to be married next year
- Own your own house this quarter
- Travel to work by car this week

Tweet - Longitudinal data enhance our ability to investigate complicated processes in the social world

Sources of Longitudinal Data

Analysing Repeated Cross-Sectional Surveys

- Often over-looked as a source of longitudinal information
- Many countries have cross-sectional surveys that are carried out on a regular basis
- They offer the possibility of pooling data for different years
- Not based on repeated contacts with the same individuals or households
- But offer opportunities to analyse general trends over time

Panel Dataset Examples (Household Panel Studies)

- US Panel Study of Income Dynamics (PSID)
  - began in 1968
    http://psidonline.isr.umich.edu/
- Germany Socio-Economic Panel (SOEP)
  - began in 1984
    http://www.diw.de/en/soep
- British Household Panel Survey BHPS
  - (1991 onwards)
  - 5k households, 10k adults, http://www.iser.essex.ac.uk/survey/bhps

http://www.understandingsociety.org.uk/

- Understanding Society (US)
  - Also known as the UK Household Longitudinal Study (UKHLS)
- Began in January 2009
- Incorporates and extends the BHPS
- 40k UK households (4K Scottish households)
- 4k households in a special ethnic minorities sample
- Innovations include:
  - Linking to administrative data; spatial data; biometric data; qualitative data; child data (from age 10)

Understanding Society Sample

- Approx. 27,000 households -
  - The fieldwork for this sample commenced in January 2009

- A boost ethnic minority sample,
  - focussed on five main ethnic minority groups, comprising 4,000 households

- Incorporating the BHPS sample of approximately 8,400 households

- An Innovation Panel of 1500 households to enable methodological research
  - (panel began in January 2008)

Understanding Society

- Focus on new research issues

- Opportunities for mixed methods:
  - Data linkage admin, organisation, spatial
  - Bio-markers and health indicators
  - Qualitative data
  - Other non-standard data: diaries, visual, audio

Administrative Data

- ONS – Longitudinal Study (England and Wales)
- Northern Ireland Longitudinal Study (NILS)
- Scottish Longitudinal Study

  A panel study of 274k people based on Census records

  http://www.lscs.ac.uk/sls/

The Workflow

The Workflow

- Planning, organising and documenting work

- This includes…

Cleaning data
Analysing data
Presenting results
Backing up and archiving material

The Workflow

Workflow should be planned and carefully orchestrated

Workflow MUST not be *adhoc*

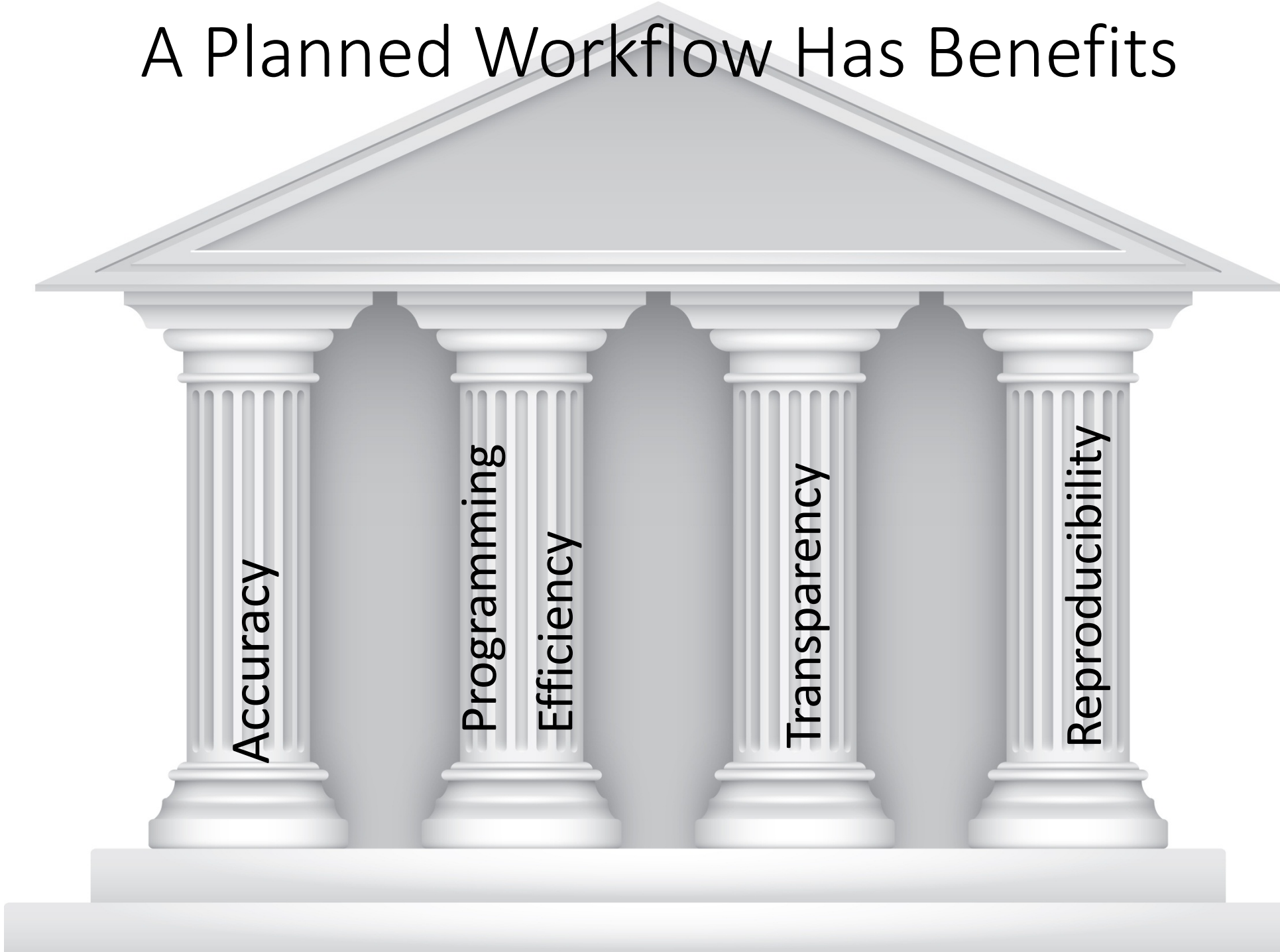
(e.g. piece-meal, developed as a reaction to mistakes etc.)

The Workflow

Better supporting YOU and what YOU DO

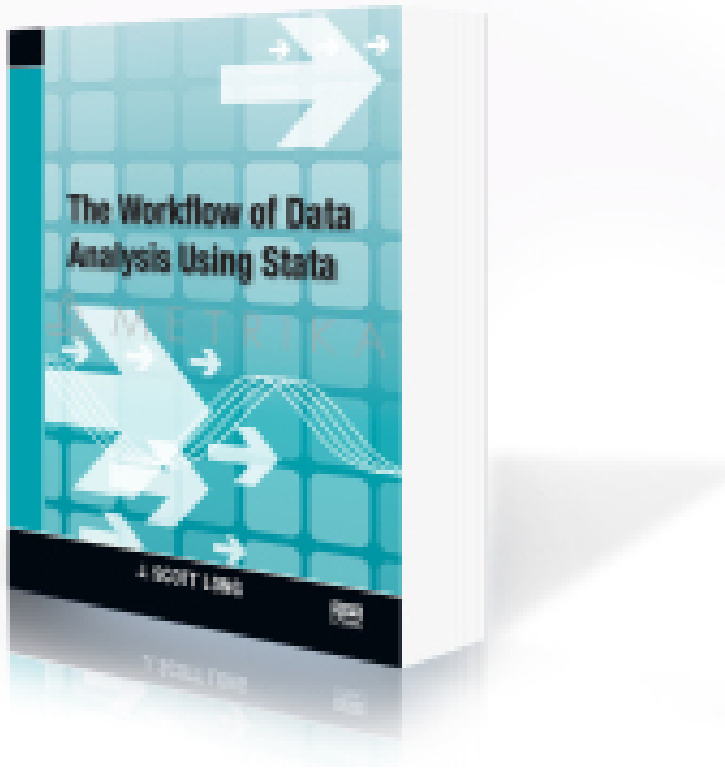
Not changing you into something YOU ARE NOT

A Planned Workflow Has Benefits



Four Pillars of Wisdom

- Accuracy
  - minimising information loss and errors in analyses and output
- Programming Efficiency
  - automation, maximising features in software
- Transparency
  - showing what you did, why, when, how
- Reproducibility
  - same results every time whoever or wherever
  - editing, rewriting reports or re-submission of papers


The Workflow of Data Analysis Using Stata
J. SCOTT LONG

The Workflow

Drukker's dictum: Never type anything that you can obtain from a saved result

My dictum (Gayle's dictum):

You can't be too fit or have too many publications

However…

Long's Law

It is always easier to document today than it is tomorrow!

Corollary 1:

Nobody likes to write documentation

Corollary 2:

Nobody every regrets having written documentation

Long's Law

Has anyone in the history of data analysis ever said

"these files are too well documented"

The Workflow

- Improving the workflow with a modest amount of effort

- The less experience you have the better
  - start from the very beginning

The Workflow

ALL SERIOUS WORK MUST BE REPRODUCIBLE!

There MUST be an audit trail

The Workflow

Why is it all so difficult?

Social science data tends to come in messy formats

Administrative data often is even more complex in nature than social survey data

The Workflow

Why is it all so difficult?

Minor decisions have major consequences…

Which cases?
Which variables?
How to code (e.g. education)?
How to recode?
Where do I truncate?

The Workflow

Minor decisions have major consequences…

Which cases?

Which variables?

How to code (e.g. education)?

How to recode?

Where do I truncate?

Can I trace these decisions in my audit trail?

The Workflow

File Naming Protocols

File Name =

name_date_depositor's initials_version_type

The Workflow

File Naming Protocols

File Name = name_date_depositor's initials_version_type

Therefore **bhpsaindresp_20140506_vg_v1.dta**

Would be a

a.. The British Household Panel Survey File “aindresp”
b.. Deposited on 6th May 2014
c.. Deposited by vg (Vernon Gayle)
d.. Version v1
e.. File type (e.g. a Stata .dta file)

	A	C	D	E	F	G	H	I
1	File Register							
2								
3								
4	Directory Name	File Name (name_subname_date[year/month/day]_depositor's initials_version_type)	File Type	Name of Author	Initials of Author	Date of Creation	Date of last revision	Brief Description of the file and it purpose
5		(e.g. bhps_aindresp_140129_vg_v1.dta)	(e.g. Stata data file)					(e.g. Stata .do file MSc dissertation; Draft Chapter 1 PhD)
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								

Other seemingly small issues such as ‘Directory Structures’ and ‘Variable Naming Conventions’ are similarly worth thinking about!

The Workflow

Why is it all so difficult?

Poor discipline and insufficient documentation

Estimating Work Time…

GOOD LUCK!

Aim for Gold in your work!

DAY 3

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh
vernon.gayle@ed.ac.uk
@profbigvern

April 2018

Analysing Repeated Cross-Sectional Data

Studying Longer Term Trends

- Many sources of ‘repeated’ cross-sectional data
- Rapid progress can be made
- Standard statistical approached (e.g. regression models)

- Comparability (equivalence) is the central challenge
- How should time be represented

Group Exercise:

Figure 1 Output: Regression model (OLS) General Certificate of Secondary Education (GCSE) points score School Year 11 Youth Cohort Study of England and Wales

Group Exercise: Interpret this model…

Number of obs = 62,910;
F(5, 62904)= 948.14; Prob > F = 0.0000;
R-squared = 0.07; Adj R-squared = 0.07

GCSE score	Coef.	Std. Err.	t	p	[95% Conf.	Interval]
cohort_1993	5.143661	.2080184	24.73	0.000	4.735944	5.551377
cohort_1995	9.365432	.2157778	43.40	0.000	8.942508	9.788357
cohort_1997	8.334607	.2166697	38.47	0.000	7.909934	8.75928
cohort_1999	13.63625	.225169	60.56	0.000	13.19492	14.07758
male	-3.119253	.1354489	-23.03	0.000	-3.384733	-2.853773
_cons	31.93301	.1675295	190.61	0.000	31.60465	32.26136

Analysing Duration Data

Alternative terminology

- Duration models
- Survival models
- Cox regression
- Cox models
- Failure time analysis
- Hazard models
- Event history analysis

Models for duration data allow the data analyst to assess the relative influence of a number of explanatory factors upon how long it takes for an event to occur

Original paper Cox (1972)

Applications

- Study the lifetimes of machine components in engineering
- Duration of unemployment in economics
- Time taken to complete cognitive tasks in psychology
- Lengths of tracks on a photographic plate in particle physics
- Survival times of patients in clinical trials

Research Examples

Heckman and Borjas (1980) used duration modelling approaches to study unemployment

Blossfeld and Hakim (1997) studied female part-time employment

Mulder and Smits (1999) investigated first time home ownership

Lillard et al. (1995) studied premarital cohabitation and subsequent marital dissolution

Research Examples

Kiernan and Mueller (1998) undertook an analysis of divorce using the BHPS and the NCDS

Boyle et al. (2008) examined union dissolution using the Austrian Family and Fertility Survey (FFS)

Chan and Halpin (2002) used BHPS to examine gender role attitudes and the domestic division of labour on divorce

Pevalin and Ermisch (2004) investigated mental health, union dissolution and re-partnering

Measuring a Duration

Three requirements for correctly determining a duration

1. A starting time must be unambiguously defined
2. Time must have a defined unit of measurement
3. The event must be clearly defined

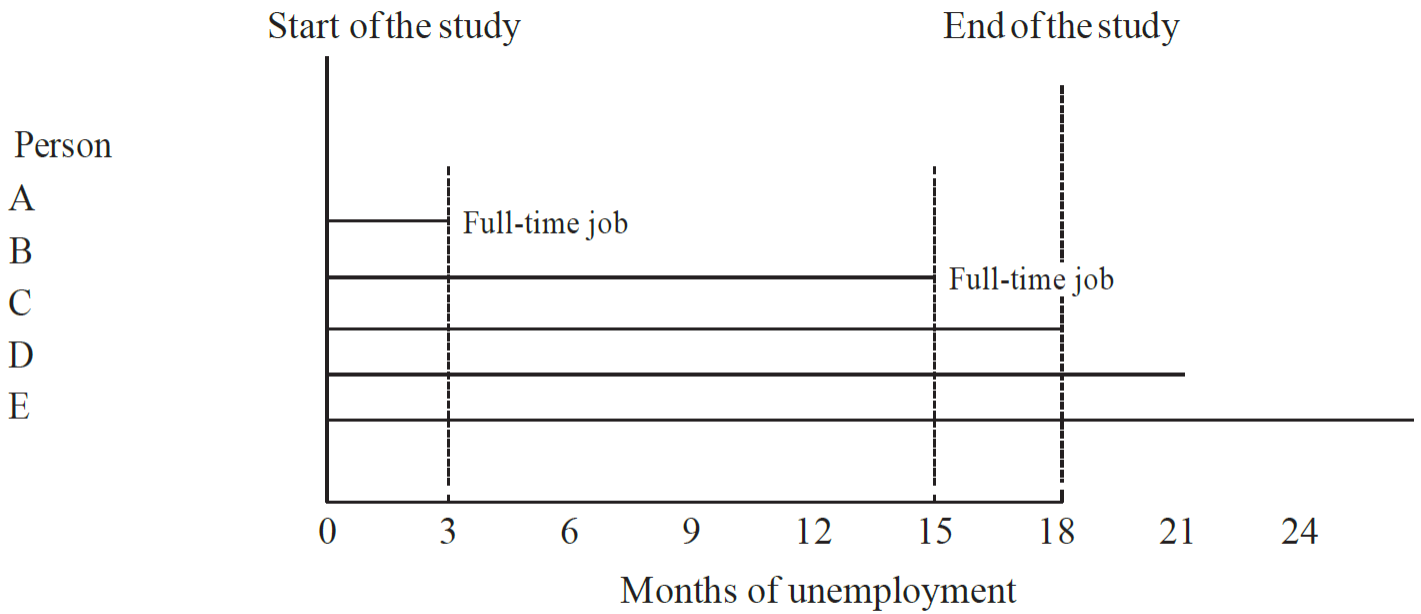


Figure 4 A diagram of a hypothetical study of unemployment

The Accelerated Life Model

- Regression models can be estimated with duration data
- Historically the log of the duration has been modelled

Censored Observations

- Censored observations affect regression model results
- The impact on the results on may sometimes be negligible
- Plewis (1997) states that when there is a very small proportion of censored cases they will have little effect, and an accelerated life model might still be suitable
- Supervisors, examiners and referees may not be convinced

Duration Modelling

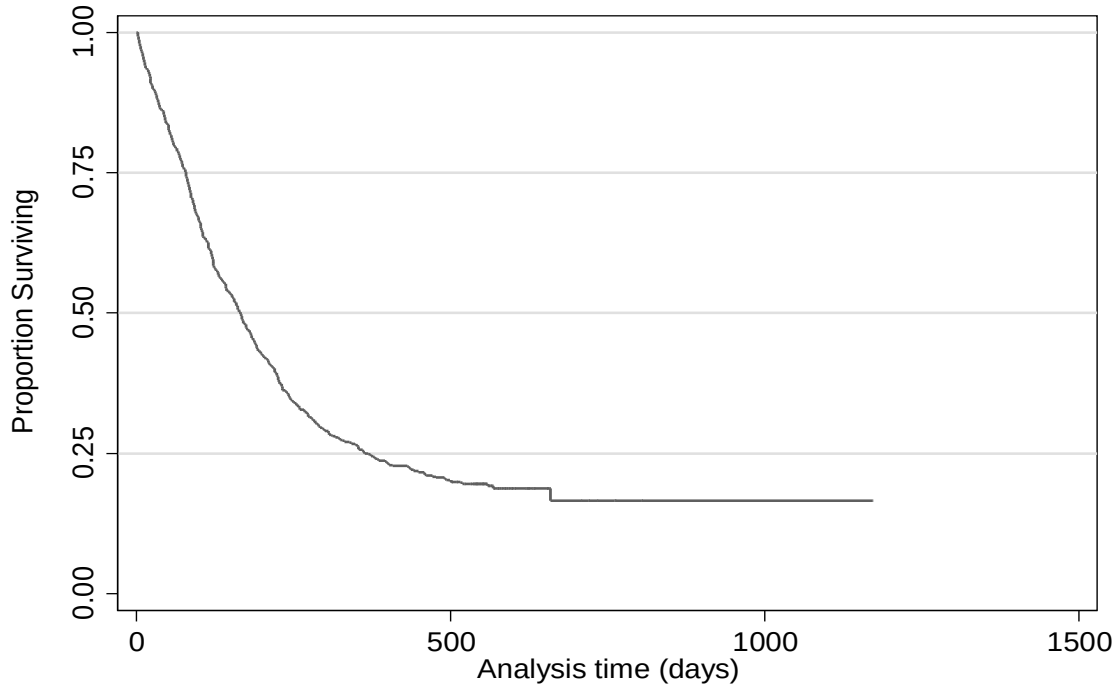
- No longer directly modelling the duration

- The focus is on modelling the probability that an event occurs at time *t* , conditional on it not having occurred before *t*

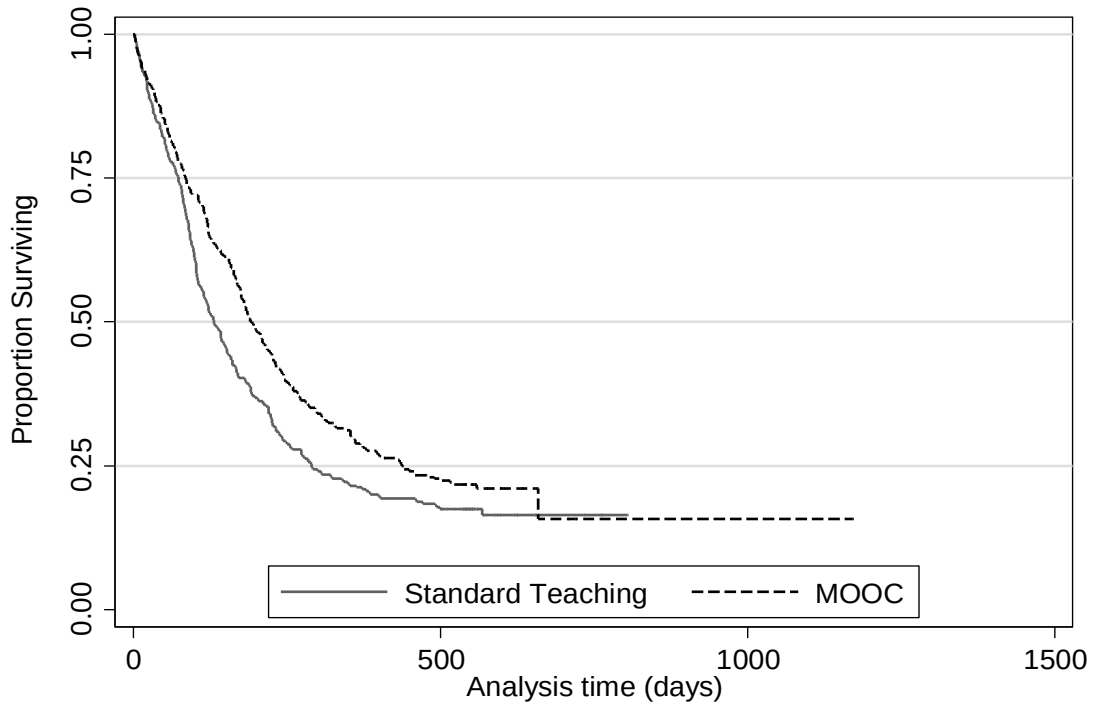
Stata Compact Codebook for the College Skills Program Dataset

Variable	Obs	Unique	Mean	Min	Max	Label
id	628	628	314.5	1	628	student id
time	628	338	234.7038	2	1172	number of days until test passed
test	628	2	.8089172	0	1	test passed (or censored)
age	623	31	32.36918	20	56	age at enrolment
no_jobs	611	28	4.574468	0	40	number of previous jobs
mooc	628	2	.4904459	0	1	taught by massive open online course
campus	628	2	.2929936	0	1	college campus
quals1	628	2	.4601911	0	1	no qualifications
quals2	628	2	.1815287	0	1	lower qualifications (below A'level)
quals3	628	2	.3582803	0	1	higher qualifications(above A'level)

Stata Output: Kaplan-Meier Plot of Time to Passing the Test (College Skills Program Data)



Stata Output: Kaplan-Meier Plot of Time to Passing the Test (College Skills Program Data)



Stata Output: Cox Regression Model Time to Passing the Test (College Skills Program Data)

```
Cox regression -- Breslow method for ties

No. of subjects =          610                 Number of obs    =         610
No. of failures =          495
Time at risk    =       142994
                                               LR chi2(6)       =       34.94
Log likelihood  =   -2851.0863                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0237543   .0075611    -3.14   0.002    -.0385737   -.0089349
     no_jobs |    .034745   .0077538     4.48   0.000     .0195478    .0499422
        mooc |  -.2540169    .091005    -2.79   0.005    -.4323834   -.0756504
      campus |  -.1723881   .1020981    -1.69   0.091    -.3724966    .0277205
      quals2 |   .2467753   .1227597     2.01   0.044     .0061706    .4873799
      quals3 |    .125668   .1030729     1.22   0.223    -.0763513    .3276873
------------------------------------------------------------------------------
```

Analysing Panel Data

‘The making of a causal inference is not a simple affair that can be reduced to a formula applied mechanically to a set of panel data on two or more variables’ (Duncan, 1972 p.36)

Wide format dataset – single survey contact

id	age	female	working_hours
001	20	0	37
002	30	1	39
003	40	1	45

Wide format dataset - repeated contacts

id	female	age_1971	age_1972	age_1973	work_hours_1971	work_hours_1972	work_hours_1973
001	0	20	21	22	37	40	35
002	1	30	31	32	39	40	35
003	1	40	41	42	45	45	15

Snapshot of long format dataset

id	year	age	hours	ln_wage
3	68	22	40	1.49
3	69	23	40	1.70
3	70	24	40	1.45

id: personal identification number
year: year of the survey
age: respondent's age in years
hours: number of hours per week normally worked in main job
ln_wage: log of weekly wages (adjusted for inflation)

Pooled Cross-Sectional Model

- Panel are pooled together and standard statistical model used (e.g. OLS)
- A good place to start to explore
- Results provide some initial information

Pooled Cross-Sectional Model

- Overall limitation of the pooled cross-sectional model is that it assumes that each observation (i.e. row within a long format dataset) is independent of other observations

Pooled Cross-Sectional Model

- With panel data we know that individual respondents contribute many times to the data (usually once per wave for many waves)

- Pooling all of the data violates the standard regression modelling assumption that each observation is independent

Pooled Cross-Sectional Model

- In practice standard errors that are too small
- Think about what this means for significance?

Robust Standard Errors

- Robust standard errors are sometimes known as Huber/White sandwich estimates of variance (see White, 1984, Huber, 1967)

id	year	age	hours	ln_wage
3	68	22	40	1.49
3	69	23	40	1.70
3	70	24	40	1.45

Collapsed dataset of the mean values

Id	year $\overline{x}$	age $\overline{x}$	hours $\overline{x}$	ln_wage $\overline{x}$
3	60	23	40	1.55

id: personal identification number

year $\overline{x}$: mean year of the survey

age $\overline{x}$: mean of respondent's age in years

hours $\overline{x}$: mean number of hours per week normally worked in main job

ln_wage $\overline{x}$: mean log of weekly wages (adjusted for inflation)

Between Effects Model

Estimates a standard cross-sectional model on the data

A regression model with *Y* the mean of the log of weekly wages (adjusted for inflation)

X vars

- mean hours per week normally worked in the respondent's main job
- mean age across the three waves of the survey

Between Effects Model

Because now there is only one row of data per respondent the problem of non-independence of observations in the original (long format) panel data is sidestepped

What might the limitation of this approach be?

The Fixed Effects Panel Model

- Concentrates on change over time within an individual respondent
- Can include explanatory variables that, for the individual respondent, change over time (e.g. age, monthly income and body mass index)
- In general cannot include explanatory variables that, for the individual respondent, are time-constant (e.g. town of birth, birth weight, father's occupation when respondent was aged 14)
- Has the potentially attractive property of providing robust estimates when observed explanatory variables are correlated with the unobserved effects

The Random Effects Panel Model

- Analyses both change within an individual respondent's outcomes, and differences between respondents' outcomes
- Can include explanatory variables that, for the individual respondent, change over time (e.g. age, monthly income and body mass index)
- Can include explanatory variables that, for the individual respondent, are time-constant (e.g. town of birth, birth weight, father's occupation when respondent was aged 14)
- Makes the assumption that observed explanatory variables are not correlated with the unobserved effects

Notation

Pooled Cross-Sectional Regression Model

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

Fixed Effects Panel Regression Model

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

Random Effects Panel Regression ('random intercepts' version)

(3) $Y_{it} = \beta_0 + + \beta_1 X_{1it} +\ldots+ \beta_k X_{kit} + \upsilon_i + \varepsilon_{it}$

Random Effects Model

```
Random-effects GLS regression                   Number of obs     =         32
Group variable: id                              Number of groups  =          8

R-sq:                                           Obs per group:
     within  = 0.8722                                         min =          4
     between = 0.2500                                         avg =        4.0
     overall = 0.8392                                         max =          4

                                                Wald chi2(4)      =     145.32
corr(u_i, X)   = 0 (assumed)                    Prob  chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |       .625   .4419417     1.41   0.157    -.2411899     1.49119
       wave2 |        .75   .5824824     1.29   0.198    -.3916445    1.891644
       wave3 |        3.5   .5824824     6.01   0.000     2.358356    4.641644
       wave4 |       6.25   .5824824    10.73   0.000     5.108356    7.391644
       _cons |     2.4375    .474224     5.14   0.000     1.508038    3.366962
-------------+----------------------------------------------------------------
     sigma_u |  .22658174
     sigma_e |  1.1649647
         rho |  .03645008   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Other Panel Models

Binary Outcomes

xtlogit	example in the .do file
xtprobit	example in the .do file
clogit	example in the .do file

Ordinal Outcomes

xtologit	random-effects ordered logistic models
xtoprobit	random-effects ordered probit models

Count Data

xtpoisson	panel data poisson models
xtnbreg	panel data negative binomial models

Dynamic Models

- Dynamic panel models extend panel models
- Appeal to the idea of using panel data to better understand ‘state dependence’
- Lagged dependent variables as X vars
- Complicated because the lagged dependent variables will themselves be influenced by unobserved effects

Dynamic Models

- Standard panel estimation procedures will be inconsistent with lagged dependent variables
- Arellano and Bond (1991) derived a suitable estimator which is available using the Stata command *xtabond*
- Stewart (2006) *redprob*

Where Do I Go Next?

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

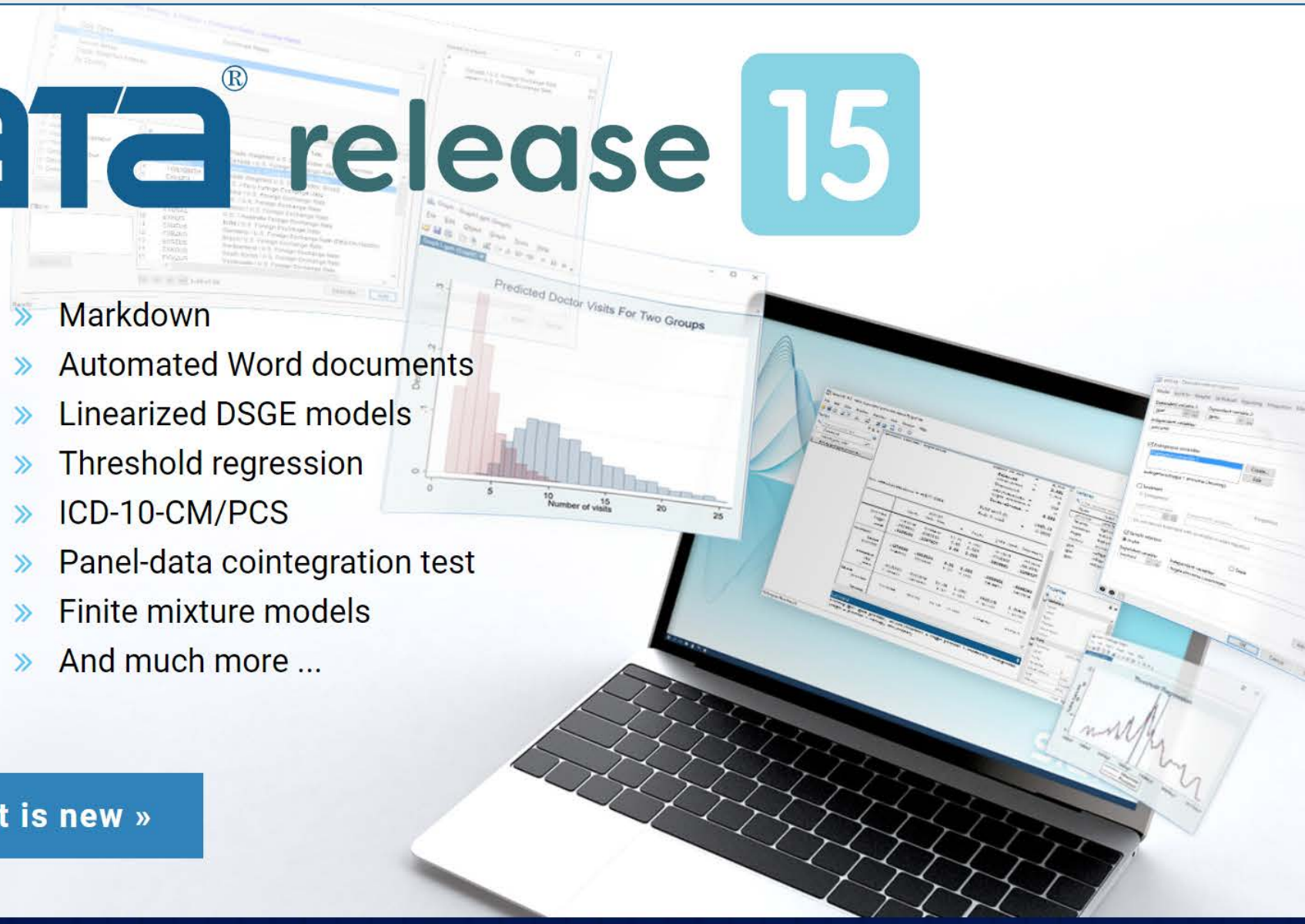
vernon.gayle@ed.ac.uk
@Profbigvern

2018

AQMEN

STATA® release 15

- Latent class analysis
- Endogeneity + Selection + Treatment
- Bayes: logistic ... and 44 more
- Nonparametric regression
- Spatial AR models
- Nonlinear multilevel models
- Mixed logit models
- Interval-censored survival
- Markdown
- Automated Word documents
- Linearized DSGE models
- Threshold regression
- ICD-10-CM/PCS
- Panel-data cointegration test
- Finite mixture models
- And much more ...



See all that is new »

Other Data Analytical Software

- SPSS https://www.ibm.com/uk-en/marketplace/spss-statistics
- R https://www.r-project.org
- SAS https://www.sas.com/en_gb/software/stat.html
- Python https://jupyter.org

Introduction

Quantitative longitudinal datasets

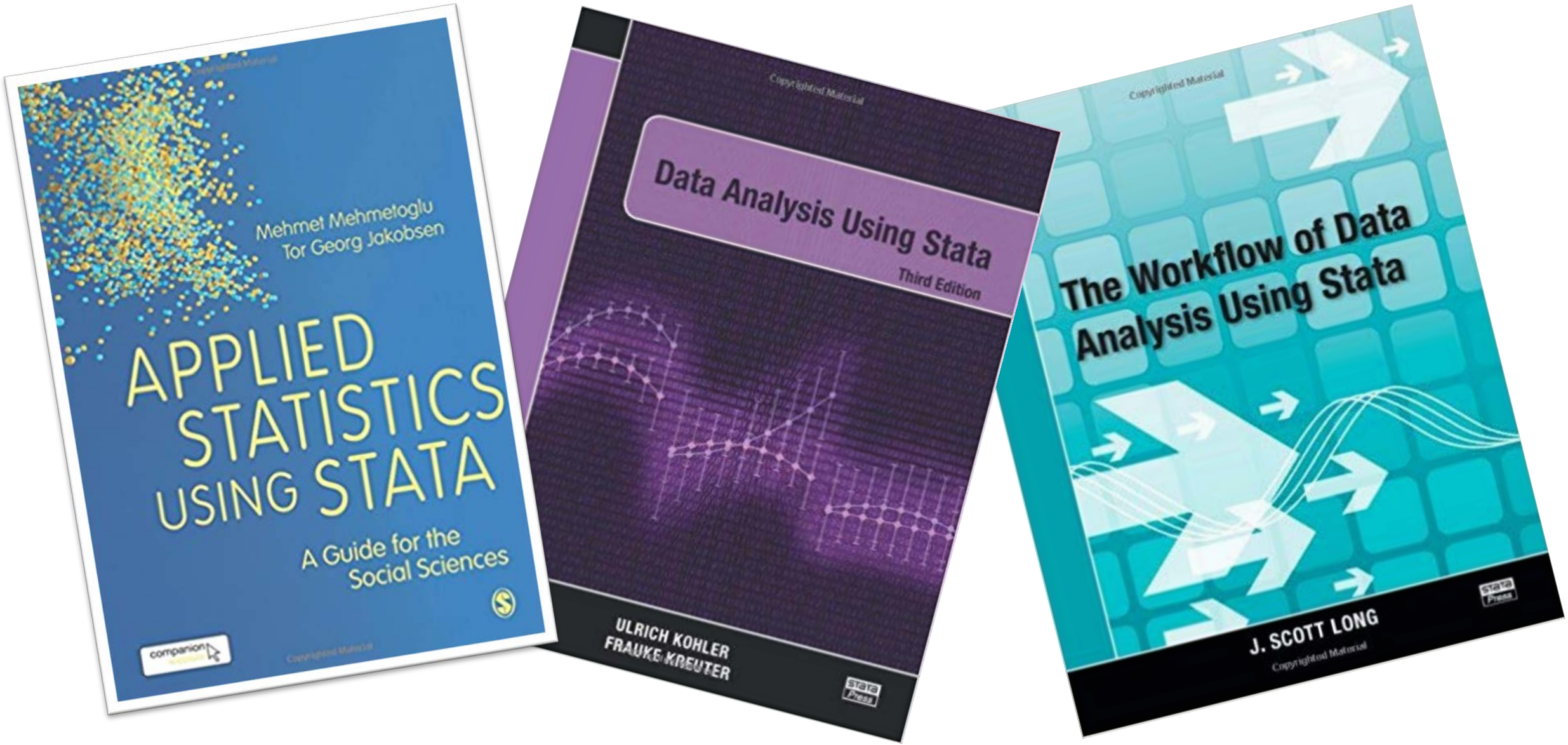
Temporal analysis with cross-sectional data

The analysis of durations

The analysis of repeated contacts data

Adopting the long view: A review of analytical methods

Getting started


Mehmet Mehmetoglu
Tor Georg Jakobsen
APPLIED STATISTICS USING STATA
A Guide for the Social Sciences
companion website
Data Analysis Using Stata
Third Edition
ULRICH KOHLER
FRAUKE KREUTER
stata Press
The Workflow of Data Analysis Using Stata
J. SCOTT LONG
stata Press


QUANTITATIVE
DATA ANALYSIS
Doing Social Research to Test Ideas
Donald J. Treiman
Multilevel and Longitudinal Modeling Using Stata
Volume II: Categorical Responses, Counts, and Survival
Third Edition
SOPHIA RABE-HESKETH
ANDERS SKRONDAL
Applied Longitudinal Data Analysis
MODELING CHANGE AND EVENT OCCURRENCE
Judith D. Singer
John B. Willett

Stata home page *http://www.stata.com/*

Stata Bookstore http://www.stata.com/bookstore/books-on-stata/

UCLA Academic Technology Services
http://www.ats.ucla.edu/stat/stata/

Princeton Stata resources http://data.princeton.edu/stata/

The website of the ESRC 'Longitudinal Data Analysis for Social Science Researchers' project much of our earlier training resources are available on this site www.longitudinal.stir.ac.uk

Stata on Twitter http://twitter.com/#!/stata

Stata Journal http://www.stata-journal.com/

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

https://github.com/vernongayle/spring_into_longitudinal_data_analysis

2018

AQMEN