

Some Uses (and Abuses!) of Regression

John Curtice
University of Strathclyde

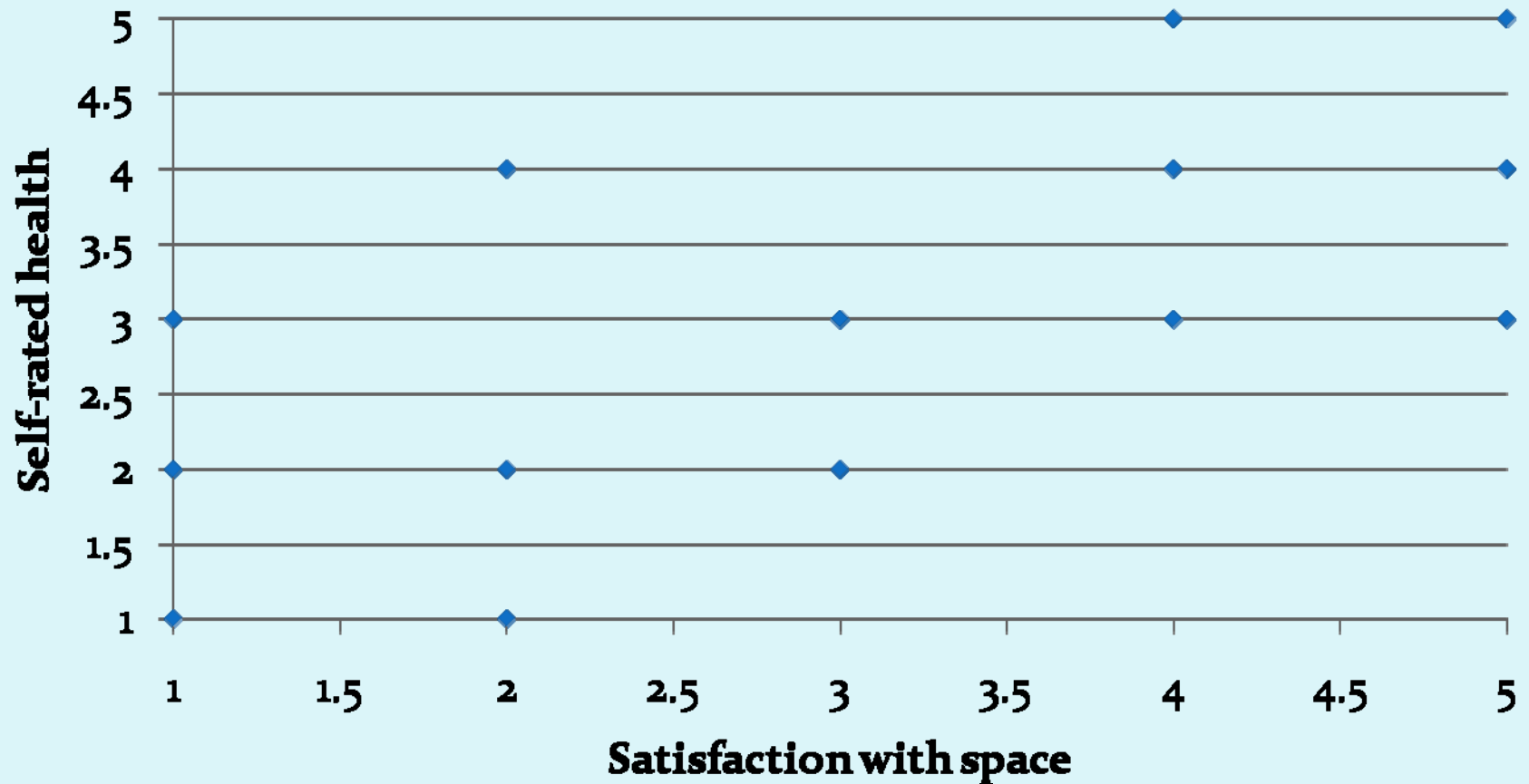
An important question

- Does the provision of a 'nice' environment help improve people's health – because it encourages them to walk, facilitates mental health, etc.?
- And so justify spending public money on provision of accessible green/open space?

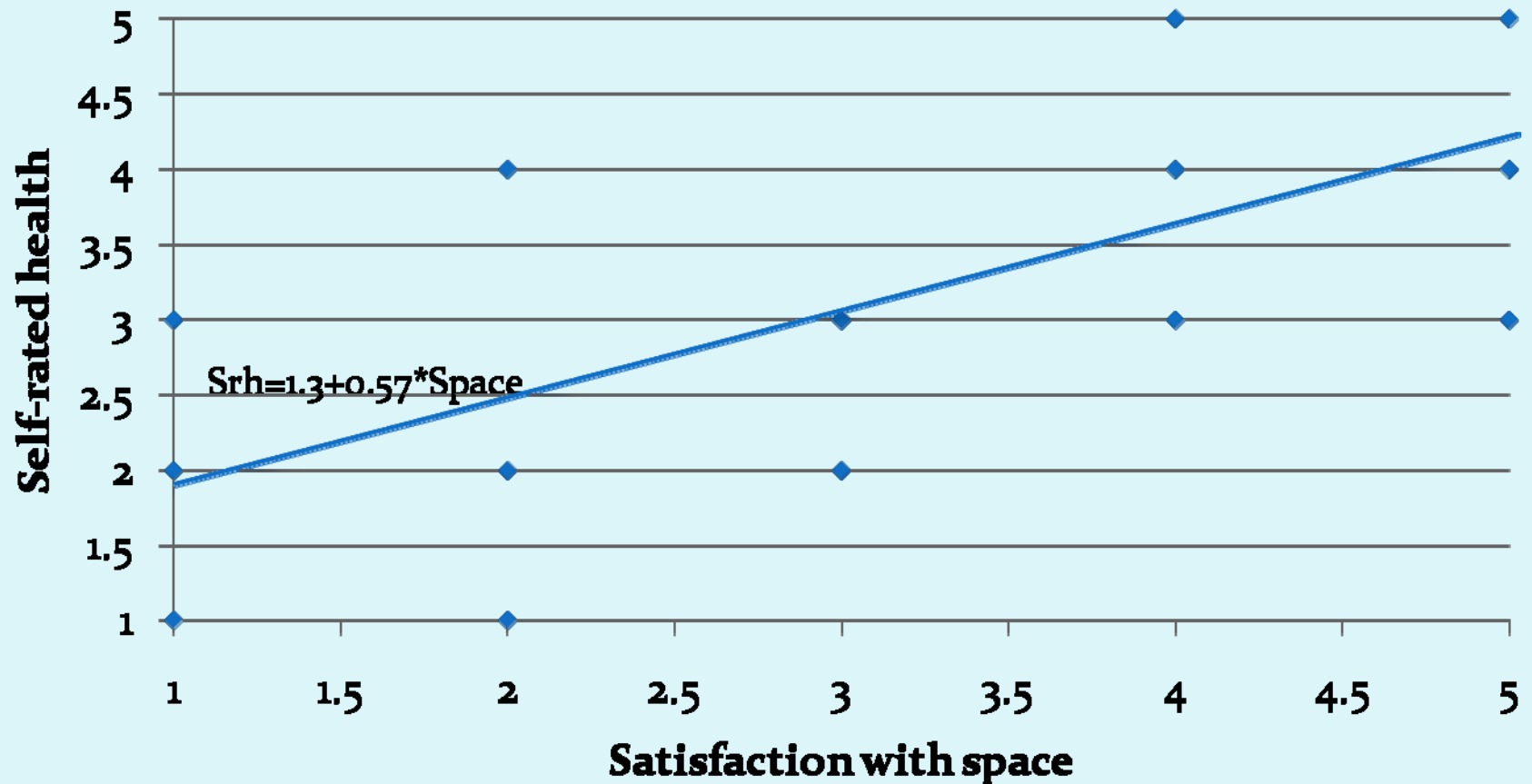
We have a survey that contains...

- Respondent's self reported health (a surrogate for actual health?)
- Respondent's reported satisfaction with the quality of the nearest accessible green/open space (such as wood, park, beach)

A look at the data



Adding a regression line



So we acquire...

- An indication of how much an increase in the independent variable (satisfaction with green space) results in better (self-reported) health
- In the form $y = a + bx$
- And an indication, using standard probability theory, as to whether 'b' (and indeed 'a') is statistically different from 0.
- And thus an evidence base for policy!

However

- There are lots of other things that might contribute to an individual's health.
- Such as age, income, marital status (depending on gender?), employment status, living in deprived area.
- And they might affect satisfaction with open/green space too (e.g. older people less likely to feel safe?)
- And once we take all this into account.....

Perhaps

- ...Satisfaction with nearest green or open space does not really make much difference to self-rated health after all.

And so...

- We construct a multivariate regression
- Of the form $y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$.
- Where y = self-rated health
- $x_1, x_2, x_3, \dots, x_n$ represent age, income, etc. including satisfaction with green space
- $b_1, b_2, b_3, \dots, b_n$ represent the estimated impact of age, income, etc. on self-rated health
- Controlling for the impact of (and inter-relationships between) all the other variables

In this case

- Age
- Income,
- Living in an area of multiple deprivation, and
- Satisfaction with local area in general
- ...are all significantly related to self-rated health
- But still so also is...
- ...satisfaction with nearest open and green space

But what about green space matters?

- Feel there is somewhere locally in which it is pleasant to walk
- Use local green space to...
- ...meet people or go with family/friends (rather than for fresh air, walk dog, etc.)
- So perhaps it is the ability of local green space to facilitate social interaction that matters?

A few technical issues - 1

- Data were clustered by postcode sector. Lots of people in some parts of Scotland interviewed, while nobody in other areas interviewed at all.
- People living in the same area may be talking about the same green space - and so give similar answers.
- And so we do not have as many independent responses as we might think – affects assessment of significance.

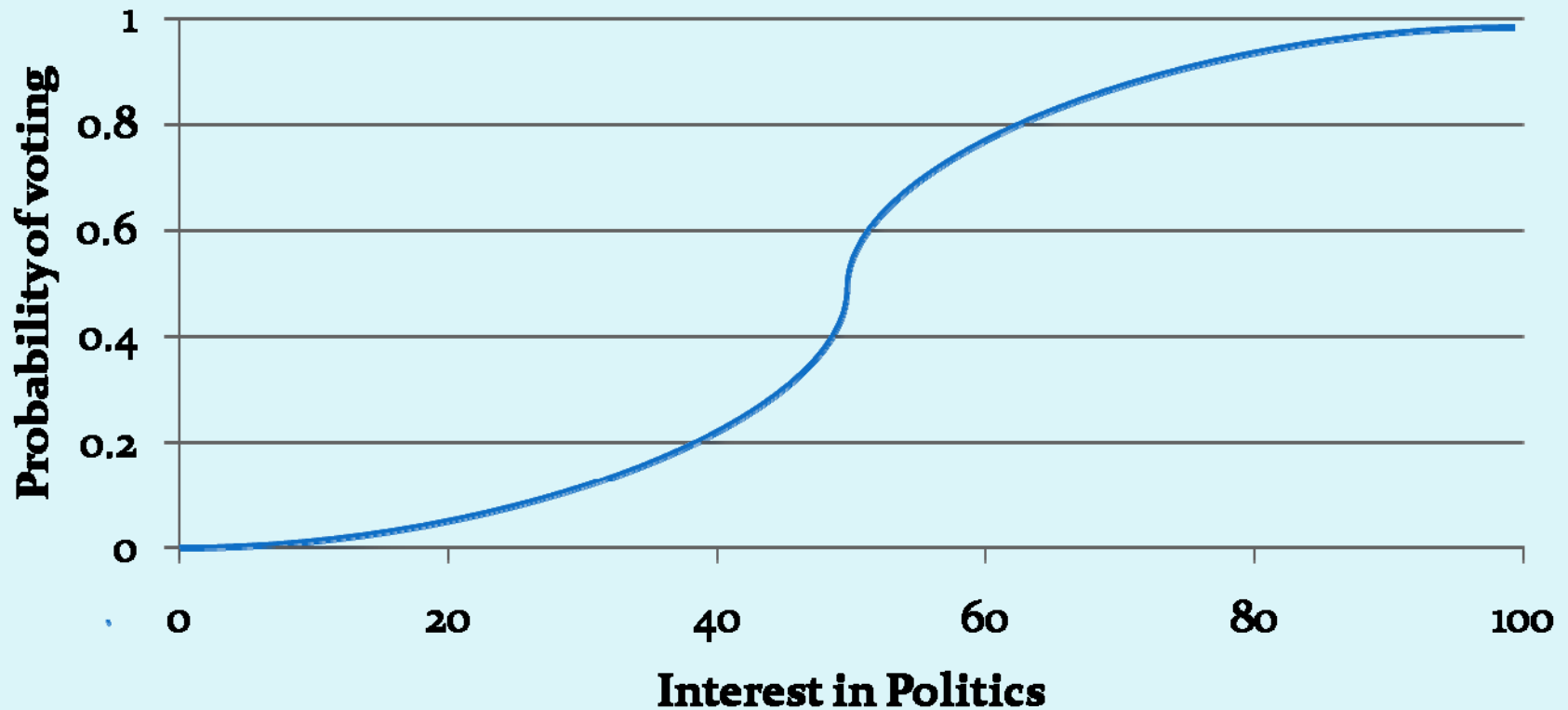
A few technical issues - 2

- Is the difference between someone in 'very good' health and someone in 'good' health necessarily the same as...
- ...the difference between someone in 'good' health and someone in 'fair' health?
- Ordinal regression drops this assumption (in the dependent variable)

A few technical issues - 3

- Life does not always go in a straight line!
- Various ways – e.g. fitting x^2 term – of adapting standard linear regression
- But still problem of binary (0,1) dependent variables – e.g. vote or not
- Chances of someone voting cannot be less than 0 or greater than 1 – so do not want predicted values outside that range.

Fit logistic/normal curve (something) like this!



And one large substantive issue

- Regression implies that changes in the independent variable cause changes in the dependent variable.
- But nothing in the technique proves that the independent variable is the cause of the dependent variable rather than vice-versa.
- Perhaps healthier people are more likely to be satisfied with their local green space because their health enables them to use it!

So always remember

- Nothing, but nothing in the technique takes away the need for prior thought and theorising about the relationships in which you are interested.
- (Or the need to be clear that your measures are valid and reliable).
- Statistical processing in the absence of thought is abuse not use!

The Dangers of Going Fishing

- Imagine a dataset in which people tell you whether they have voted or not in a recent election
- And lots and lots of other variables including....

This Lot!

Demographics	Motivations/Behaviour	Environment
Age	Strong attachment to party	Trust Politicians
Gender	Feel duty to vote	Devolution effective
Social Class	Interested in politics	Distance to Polling Station
Education	Feel voting makes a difference	Marginal Seat
Housing Tenure	Voted previous election	
Economic activity	Belong to voluntary organisation	
Marital status	Attitudes towards nuclear power	

You threw it all in

- And find that the following are all significant...
- Nearly all the demographics
- Most of the individual attitudes & behaviours – including voted last time and attitudes towards nuclear power
- Distance to polling station & trust in politicians
- In short, much like Britten's Old Joe, you have found yourself 'a shoal'.

And so you conclude

- The reasons why people vote are ‘complex’ (well, thanks a bunch)
- Including...
- That people who voted were also more likely to have voted at the previous election (highly illuminating)
- Attitudes towards nuclear power – but are at a loss to explain why (so that’s clarified things)

Find the model that addresses the question

- Debate about the role of choice in the provision of public services
- Protagonists argue...
- Both that choice generates competition which in turn results in more effective, user-orientated services
- And that users value choice for its own sake.

So how can we tell...

- ...whether choice is valued for its own sake?
- If we show that those who think they have a sufficient choice of hospital have higher levels of satisfaction with inpatient services
- This might be because the provision of choice results in better services
- Or because choice is valued for its own sake

An answer?

- Also include in a multivariate regression perceptions of how good/bad various specific aspects of inpatient services are.
- Hospital doctors take complaints seriously
- Nurses take complaints seriously
- Doctors tell you what you need to know
- Doctors take patient views seriously
- Operations take place on day booked
- Only allowed home when well enough

Two models of satisfaction with inpatient services

Independent Variables	Model A	Model B
Perceived Choice	0.32**	0.08
Perceived Performance	-	1.00**
Age 65+	0.85**	0.69**
Recent inpatient	0.38**	0.41**
Conservative supporter	-0.33*	-0.20

So...

- Perhaps choice is not valued for its own sake after all
- In which case the argument for introducing choice rests on whether it is an effective way of improving performance...
- ...and perhaps there are other ways of doing so ???

Conclusion

- Regression is the work horse of much quantitative social science research
- Thanks to recent statistical developments it can now be applied to more types of data
- It can be invaluable in helping us sort out the wheat from the chaff
- But only if applied with thought and care – not as a fishing net.