INTERMEDIATE QUANTITATIVE METHODS

Avoiding Pitfalls in Regression Analysis

Per Engzell (Nuffield College)

Hilary term 2018

Sociology Department, University of Oxford

REGRESSION ARTEFACTS

PRELIMINARIES

Last week we learned that selection on y will often lead to underestimates, while selection on x will have little consequence

Are there any exceptions to this? You guessed it… yes there are

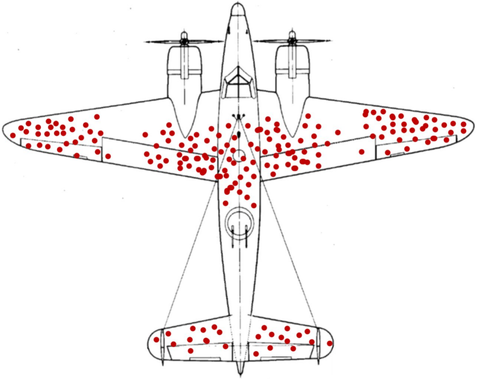
Regression artefacts can come in a number of different forms, but we are going to look at three of them, all related to selection

- Survivorship bias
- Regression to the mean
- Lack of common support

These are not the only ways your data can fool you, but they are some of the more common ways, and knowing them will be helpful

SURVIVORSHIP BIAS

Problem: Conditioning on ”survival” in some sense often introduces spurious relationships



SURVIVORSHIP BIAS

Often this can be rephrased as a problem of controlling for a variable that should not be controlled for or ”conditioning on a collider”

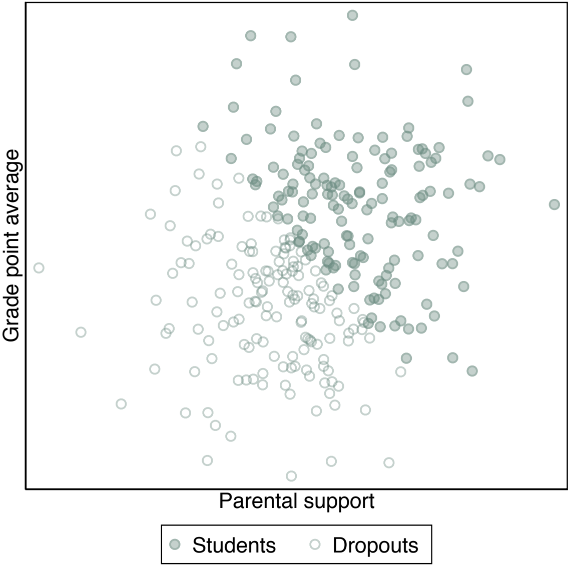
Selection on a variable y that is caused by two variables x_1 and x_2 alters the relationship between x_1 and x_2

Example: we want to know whether economic support from parents increases or decreases the motivation to study

We collect data on university students and look at whether the receipt of parental support is related to grades

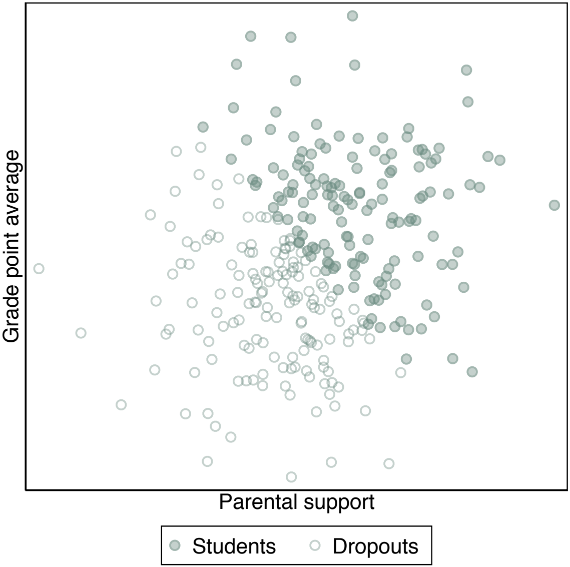
SURVIVORSHIP BIAS

- Some students drop out during their studies, but we only observe those who remain



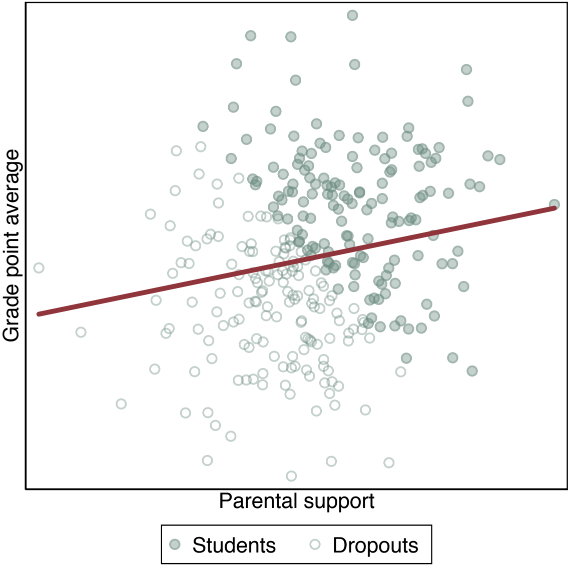
SURVIVORSHIP BIAS

- Some students drop out during their studies, but we only observe those who remain
- Both parental support and high grades help students stay



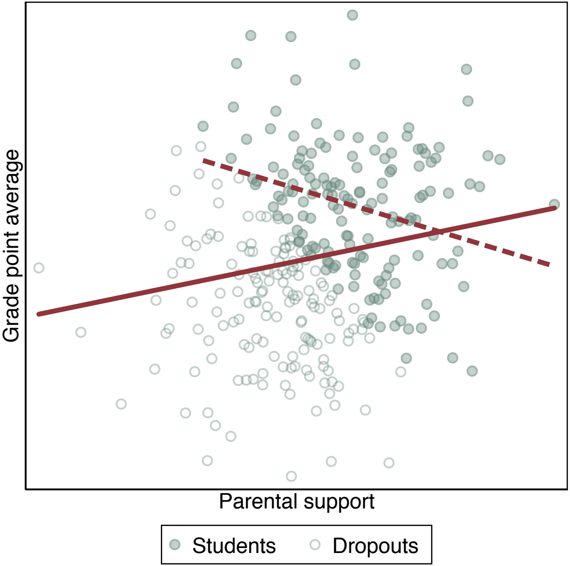
SURVIVORSHIP BIAS

- Some students drop out during their studies, but we only observe those who remain
- Both parental support and high grades help students stay
- Corr (all): +0.19



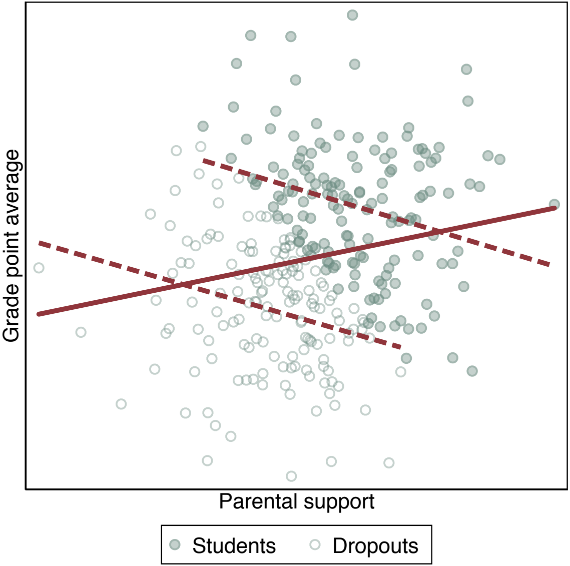
SURVIVORSHIP BIAS

- Some students drop out during their studies, but we only observe those who remain
- Both parental support and high grades help students stay
- Corr (all): $+0.19$
- Students: -0.28



SURVIVORSHIP BIAS

- Some students drop out during their studies, but we only observe those who remain
- Both parental support and high grades help students stay
- Corr (all): $+0.19$
- Students: -0.28
- Dropouts: -0.28



REGRESSION TO THE MEAN

Problem: Conditioning on a random variable x at time t will tend to induce a less extreme value at $t + 1$

Illustration:

A psychologist is lecturing a group of military instructors, and tells them that rewards are usually better than punishment at changing people's behaviour

Someone in the audience protests: ”I have often praised a cadet when they perform a maneuvre well and always find that they do worse on the next try. But when I have punished them for a bad performance, they always do better the next time!”

If you were the psychologist, should this change your view?

For more on this anecdote, see Daniel Kahneman: Thinking, Fast and Slow (Penguin Books, 2011).

REGRESSION TO THE MEAN

With random variation, those with an unusually high or low value of y at time t are likely to show a less extreme value at $t + 1$ (the same is true for any other time point than t, e.g. $t - 1$)

If we are selecting on the value at time t, we risk interpreting subsequent regression to the mean as a meaningful change when all we are seeing is random variation

A slightly more involved example

We want to know how children with high and low ability develop depending on their social environment (high or low SES)

Based on an initial test score, we divide children into four groups – high ability–low SES; high ability–high SES; low ability–high SES; low ability–low SES – and follow them with repeated tests

REGRESSION TO THE MEAN

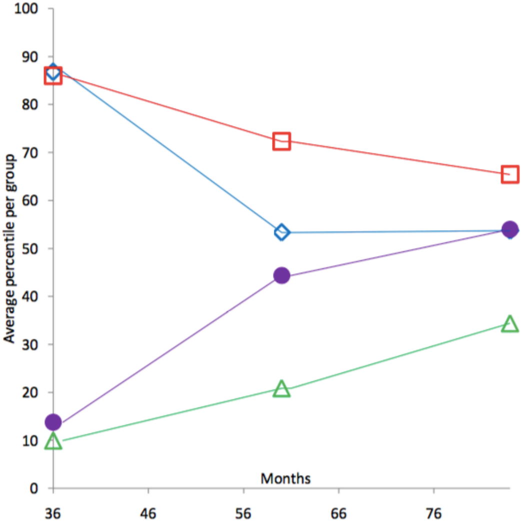
- What do you see?

high ability–high SES

high ability–low SES

low ability–high SES

low ability–low SES



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

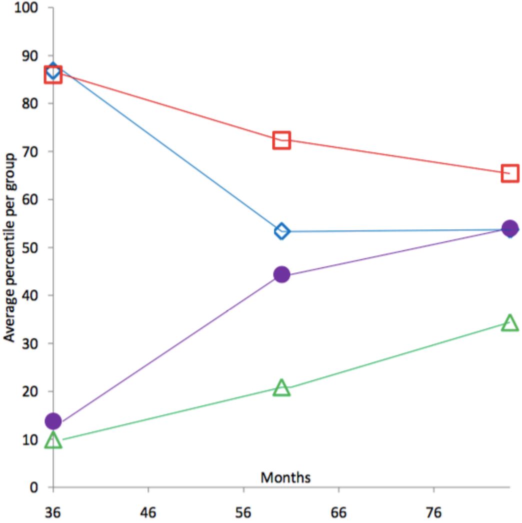
- What do you see?
- Initially high achieving children from low-SES homes fall behind

high ability–high SES

high ability–low SES

low ability–high SES

low ability–low SES



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

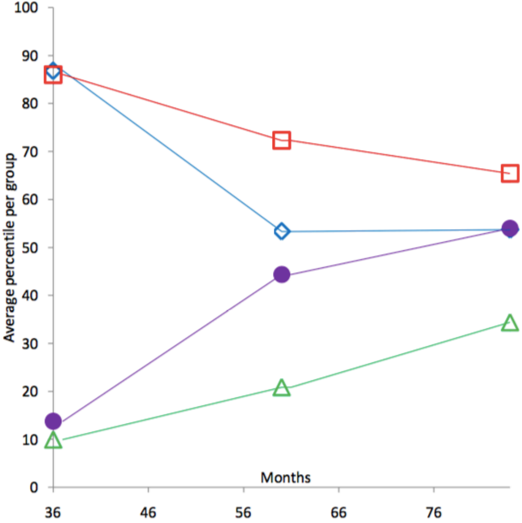
- What do you see?
- Initially high achieving children from low-SES homes fall behind
- Initially low achieving children from high-SES homes overtake them

high ability–high SES

high ability–low SES

low ability–high SES

low ability–low SES



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

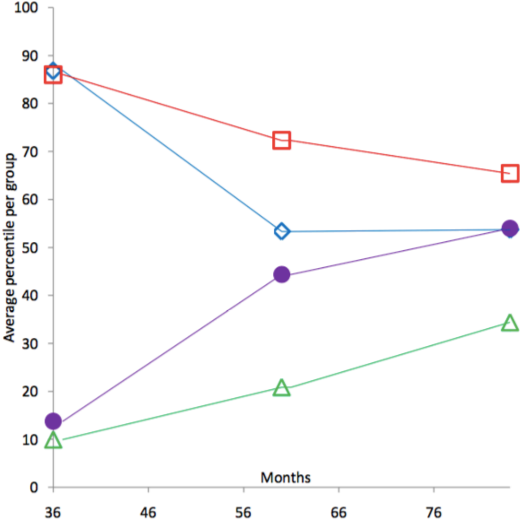
- What do you see?
- Initially high achieving children from low-SES homes fall behind
- Initially low achieving children from high-SES homes overtake them
- This looks bad!

high ability–high SES

high ability–low SES

low ability–high SES

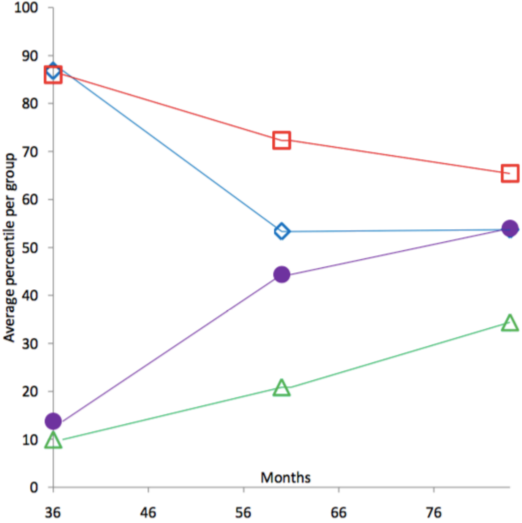
low ability–low SES



Source: Jerrim, John, and Anna Vignoles. "Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes." Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

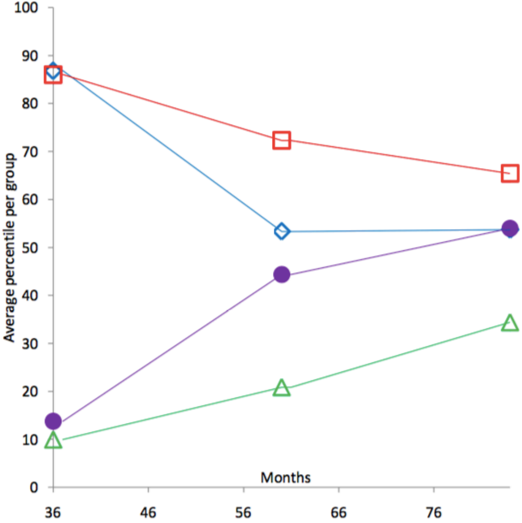
- However, some of those with top scores are likely to have had a lucky day



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

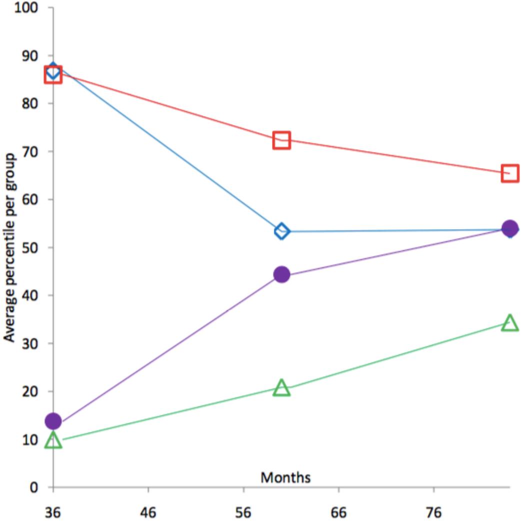
- However, some of those with top scores are likely to have had a lucky day
- Especially true in low-SES homes, where average achievement is lower



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

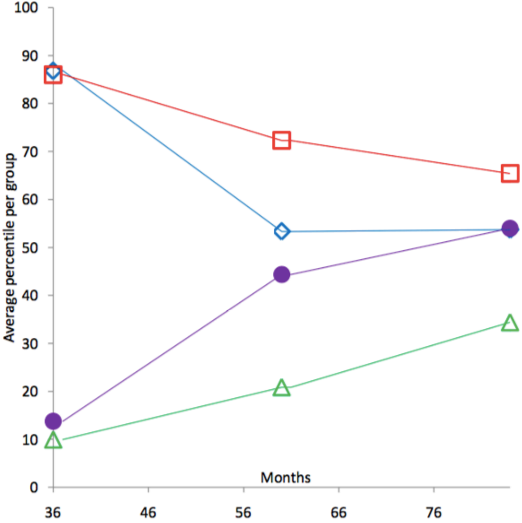
- However, some of those with top scores are likely to have had a lucky day
- Especially true in low-SES homes, where average achievement is lower
- The opposite holds for the bottom of the distribution



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

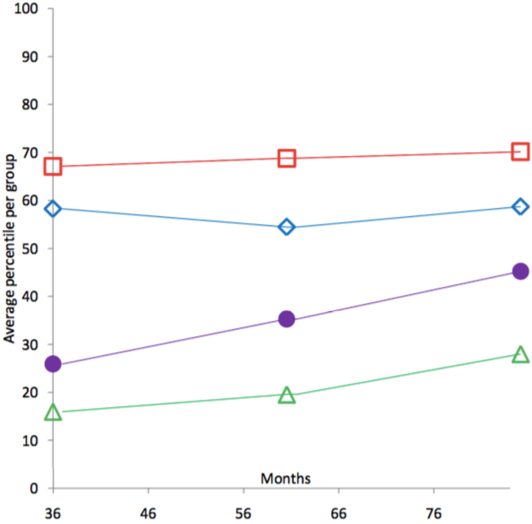
- However, some of those with top scores are likely to have had a lucky day
- Especially true in low-SES homes, where average achievement is lower
- The opposite holds for the bottom of the distribution
- Using a different test taken at t1 to determine the groups changes this



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

REGRESSION TO THE MEAN

- However, some of those with top scores are likely to have had a lucky day
- Especially true in low-SES homes, where average achievement is lower
- The opposite holds for the bottom of the distribution
- Using a different test taken at t1 to determine the groups changes this



Source: Jerrim, John, and Anna Vignoles. ”Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 4 (2013): 887-906.

LACK OF COMMON SUPPORT

Problem: Extrapolation beyond the range of available data makes unrealistic demands on having the model correct

LACK OF COMMON SUPPORT

Suppose we wanted to know whether children of low socioeconomic origin (SES) do better in schools that are Catholic

We have data where:

- GPA = student's grade point average (mean 0, s.d. 1)
- SES = a standardised measure of social origin (mean 0, s.d. 1)
- Catholic = 1 if the school a student goes to is Catholic, 0 otherwise

We estimate the regression

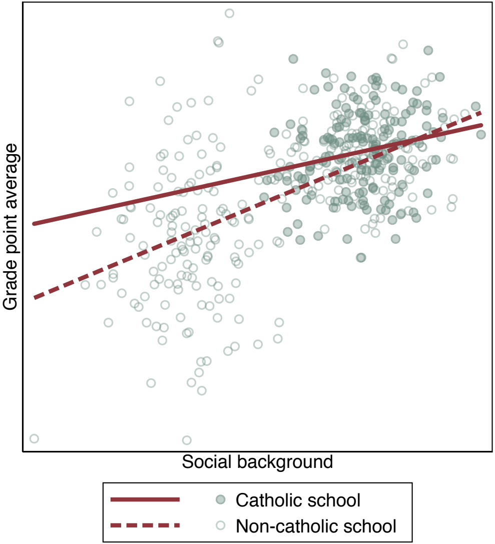
$$\text{GPA} = \beta_1 + \beta_2\text{SES} + \beta_3\text{Catholic} + \beta_4(\text{SES} * \text{Catholic}) + \text{u}$$

Our results tell us that $\beta_2 = 0.45$, $\beta_3 = 0.25$, $\beta_4 = -0.20$

How do you interpret this?

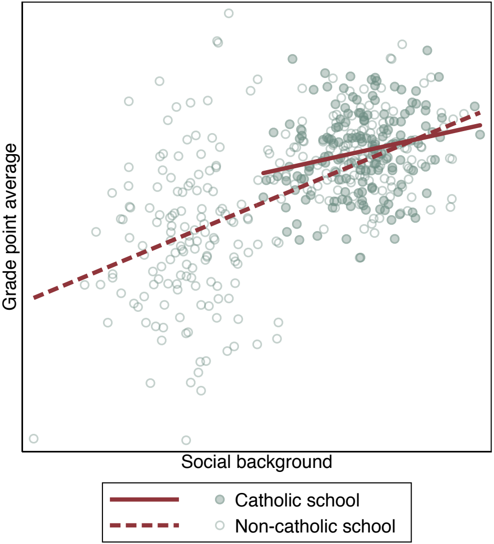
LACK OF COMMON SUPPORT

- Fitting a linear regression seems to show that low-SES children do better in Catholic schools



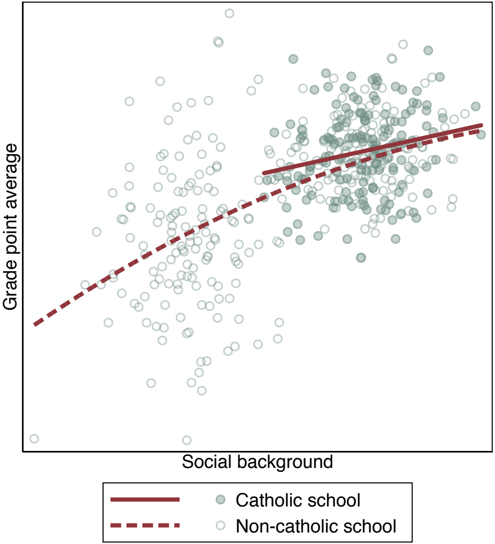
LACK OF COMMON SUPPORT

- Fitting a linear regression seems to show that low-SES children do better in Catholic schools
- But: the range of SES is also more limited there



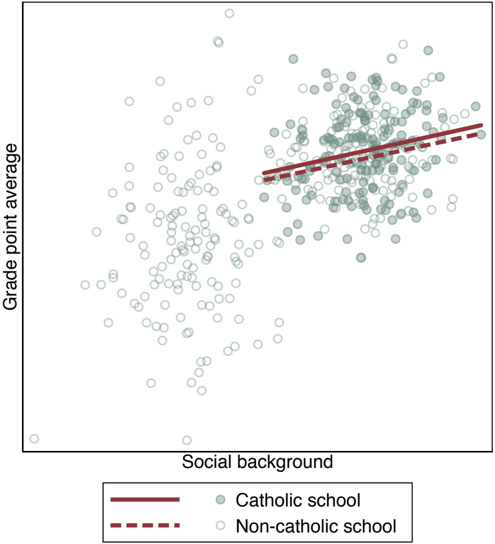
LACK OF COMMON SUPPORT

- Fitting a linear regression seems to show that low-SES children do better in Catholic schools
- But: the range of SES is also more limited there
- At higher levels of SES the relationship looks more similar



LACK OF COMMON SUPPORT

- Fitting a linear regression seems to show that low-SES children do better in Catholic schools
- But: the range of SES is also more limited there
- At higher levels of SES the relationship looks more similar
- Comparing only across range of common support, interaction disappears



STATISTICAL ARTEFACTS

What to do? Some correction methods

- Heckman correction addresses (one type of) survivorship bias
- For regression to the mean, you have seen one solution
- Matching methods often used for lack of common support

Instrumental variables is a general solution to endogeneity, but depends on having a good instrument

These are methods that work for specific problems, or with specific data, and under strong assumptions. More generally, there is no insurance

STATISTICAL ARTEFACTS

That being said...

Knowing your data helps. Have a good grasp of distributions and correlations before you go into more complex relationships. Visualise early, visualise often

Generally, the more complexity you introduce into your models, the more can go wrong. Do the simple first and increase complexity gradually

”Begin with very simple methods. If possible, end with simple methods.” – Sir David Cox

WHAT HAVE WE LEARNED?

“Pseudo-facts have a way of inducing pseudo-problems, which cannot be solved because matters are not as they purport to be.”

- Robert K. Merton, “Notes on Problem-Finding in Sociology,” Sociology Today, Robert K. Merton, et al. (eds.). Basic Books (New York, 1959), p. xv.

“Before you come up with some smart explanation of how the pig got into the tree, just be sure that it is the pig that is in the tree.”

- William H. Sewell, quoted by John H. Goldthorpe, Sociology as a Population Science. Cambridge University Press (Cambridge, 2015), p. 87.

Questions?