Regression and dependence

...

PLSC 309
20 March 2019

Review: statistical modelling

- A statistical model finds a functional relationship between explanatory variables (X) and an outcome (Y)
- We do this so we can predict new values for Y based on a hypothetical set of X
- It also helps us learn relationships between X and Y
    - Based on interpreting model parameters

Review: linear regression

- Linear regression is the most basic statistical model
- It finds the best straight line relationships between X and Y
- Requires the effect of X on Y to be additive
- If the relationship between X and Y is additive, our residuals will be
    - Homoscedastic (constant variance)
    - Independent
    - Normally distributed

Review: OLS

- Ordinary least squares finds the best fitting line by minimizing the sum of squared errors, or residuals
  - $\hat{Y}$-Y
- Because linear models are additive, we can find this bet fit by working with one X variable at a time

Review: OLS estimator for β

- The question is, how do we find the best slope

$$b_1 = \frac{s_y}{s_x} R$$

- The correlation (strength of linear relationship)...
- ... multiplied by the change in Y over the change in X

Assumptions

We make one large, overarching assumption: that our model is additive

We can see if this assumption holds by looking at our residuals. Are they...

- Homoscedastic? (have constant variance or spread)
- Normally distributed?
- Related to any of our explanatory variables?

Additivity, linearity, and normality

- You've seen these terms come up a lot, often in similar contexts. How are they related?
- If a function is additive it is necessarily linear
  - The effect of X on Y can be summarized with a slope and y-intercept
- If a function is linear, it is necessarily normal
  - Remember the CLT: all sums are normal
  - Linear functions are additive, therefore they are a type of sum
  - Therefore they are normal

Additivity comes from the central limit theorem (CLT)

- Additivity assumes we can take explanatory variables (X) in and out of our model, and they won't change the estimates for the slope of other variables
- If the value of the slope for one variable doesn't change when the other is eliminated...
- ...these two variables are independent from one another
- In other words, if variables are independent, then they are additive

Independence between two explanatory variables

- Two variables, X_1 and X_2 are independent if they do not affect one another
  - E.g. the value of one is *not dependent* on the value of the other
- Formally, we can write this as $P(X_1|X_2) = 0$ or $P(X_1 \text{ and } X_2) = P(X_1) + P(X_2)$
- Note that X_2 or X_1 do not have to be in the model to violate independence
  - When they are both in the model and dependent, this violates additivity
  - When one is absent from the model, there is omitted variable bias

Intuition behind independence

If knowing the value of one variable allows you to better guess the value of your given variable, there is a violation of independence

Dependent: Measuring your weight across every day for 100 days

Independent: Taking a random sample of 100 people and measuring their weight

Dependent data

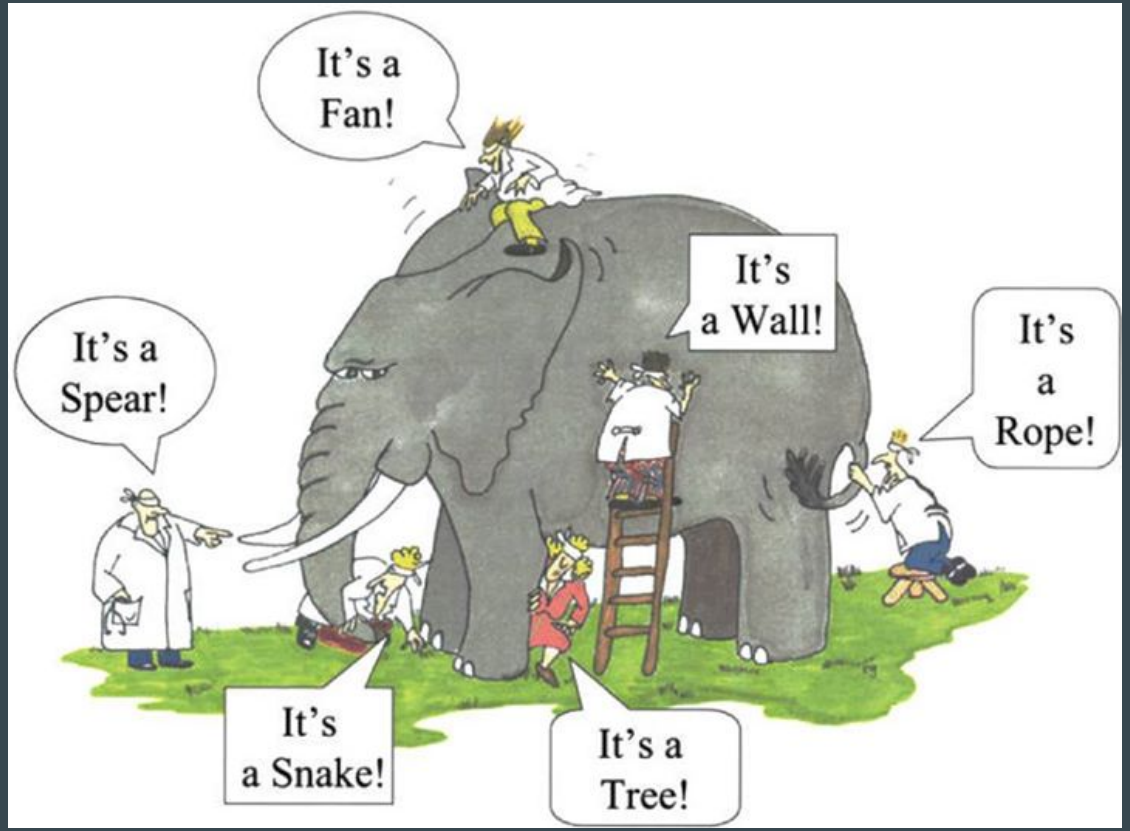
Social science is full of dependent data!

- Many concepts are interrelated
  - E.g. income is related to education, political beliefs, voter turnout, etc.
- We are often interested in outcomes over time
  - E.g. economic growth, potential of political conflict, presidential approval rating
- Many outcomes are spatially clustered
  - E.g. poverty, crime, industrialization, etc.

Dependence and information

- Information is a mathematical term used to define how much our data can tell us about the real world
- In a random sample, *all observations have an equal amount of information*
  - Say we have an existing value x_i and we are looking at another value x_{i+1}
  - Since x_i tells us nothing about x_{i+1}, when we get x_{i+1} we have doubled our information!
- With dependent data, *we have less information*
  - Now, since x_i already tells us something about x_{i+1}, when we get x_{i+1} we have less total information then the first case

Dependence, information, blind men, elephants



Dependence and information example

Say you are interested in how Common Core curriculum has impacted math scores

- Independent (more information): Take a random sample of 50 students from all classes using this curriculum in the country
- Dependent (less information): Take a single class of 50 students using this curriculum

Why does dependent data mess up our models?

- A linear model assumes additivity/independence
- We calculate our errors with the formula σ/sqrt(n)
- This is based on the assumption of independence / full information
- In reality we have less information than our n would have you believe
- Therefore, our standard errors are too small

Consequences of dependence

- We just saw that having dependent data is basically like having a smaller sample size
- Since n is the denominator in the standard error equation, our standard errors will be smaller than they should if we wrongly assume independence
- If our standard errors are smaller, *then our p-values will be smaller and our confidence intervals more narrow*
- This will lead us to wrongly reject H_0 when we shouldn't!

Fixing

- Statistical models, like all computer programs, are stupid but powerful
  - They know nothing about the variables you put-in
  - It all looks like numbers on a spreadsheet
- We can fix dependence by *explicitly modelling the dependent relationship*

Common types of dependence

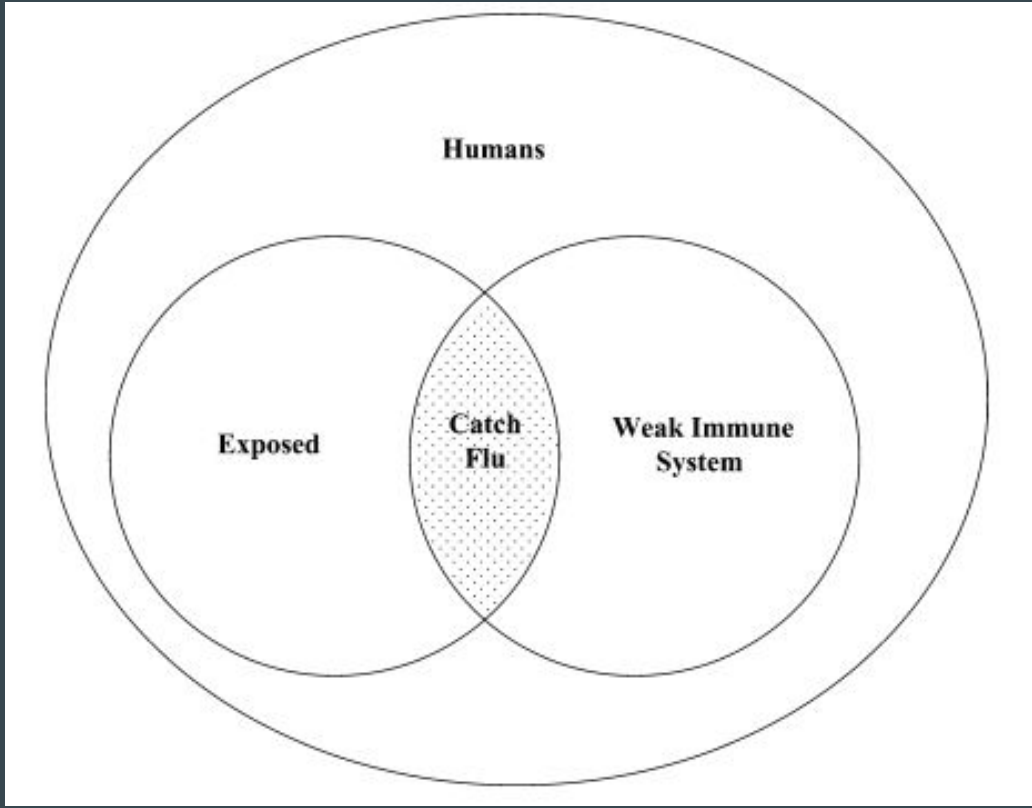
1. Interactive effects
    a. When two variables in the model are dependent on one another
2. Temporal dependence
    a. When the same variable is measured over time
3. Spatial dependence
    a. When geographic impacts the values of a variable

Interactions

An interaction between two variables means that there is some effect of these variables that only exists *when they are combined*

- Interactions can be identified as “and” relationships
- E.g. a plant needs *water and sun* to grow
- This is the logic of conditional probability

Interactions



Political interactions

Examples of interactions in political science:

1. Countries are more likely to become democratic after they have been independent longer and increase their GDP
2. A candidate will get a greater vote share if most voters rate them high on the economy and there is high unemployment
3. Poverty is associated with low voter turnout particularly for single-parent households

Modelling interactions

Interactions can be transformed into additive relationships, *as long as we include all constituent terms*

$$Y = \beta_1 X_1 + \beta_2 X_x + \beta_3 (X_1 * X_2)$$

- This divides up the relationship between X_1, X_2, and Y into two parts
- The independent relationship of X_1 and X_2 (β_1 and β_2)
- The dependent relationship of X_1 and X_2 (β_3)

Example of interactions in regression

Table 4 The determinants of multipartism with dichotomous and continuous predictor variables

Predictor variables	*Model 1* *Dichotomous variables*	*Model 2* *Continuous variables*
Electoral system permissiveness	0.55 (0.50)	−0.19 (0.30)
Social heterogeneity	−0.83 (0.55)	−0.36 (0.35)
Electoral system permissiveness × Social heterogeneity	1.65** (0.70)	0.48** (0.18)
Constant	2.52*** (0.41)	2.67*** (0.61)
N	54	54
Adjusted R^2	0.26	0.32

Temporal dependence

- Temporal dependence is a way of saying that previous values of your data influence current values of your data
- This is a problem of *time-series analysis* where we want to measure changes over time
    - GDP growth
    - Political conflict incidence
    - Candidate approval rating
    - Respect for human rights

How to deal with temporal dependence?

- For dependence between variables, we explicitly included the interaction between the two variables
- We can use a similar approach to fixing temporal dependence, by explicitly including the dependent values in our regression model

Lagged variables

- In the simplest model of temporal dependence, there effect of time is constant
  - Imagine you were trying to predict your weight, but you knew that you always gained/lost the same amount each month
- If this is true, your current value depends on the value in the previous time period
  - We call the period of time (t)
  - We don't need all periods of time, since the effect is constant
- Therefore, we just need to include the previous time period in our model

Lagged variables (equation)

Say we are interested in modelling an outcome Y that is measured over discrete periods of times (think a measure for every day, month, year, minute, etc.)

$$Y = \beta_1 Y_{t-1} + \beta_2 X_1$$

- Because it is taken from the time period before, this is known as a *lagged variable*

Lagged variables, example

- Predicting stock market price by the previous year's results

	Coeff	t Stat	P-value	Lower 95%	Upper 95%
Inter.	-0.003	-0.662	0.508	-0.013	0.006
X_{t-12}	0.022	4.833	1.5E-6	0.013	0.032

Spatial dependence

- Spatial dependence means that there are clusters in your data
- Within these clusters, knowing one point allows you to know more information than you should about another point within that closer
- In this way, space is the similar to time for temporal dependence

Spatial dependence in political science

- Any time you are dealing with meaningful geographic clusters:
  - GDP growth (western European countries share a similar set of historical factors that has produced higher GDP gains)
  - Voter turnout (states have very different standard for polling accessibility)
  - Public policy (different states have different historical legacies which can modify how policies are implemented)

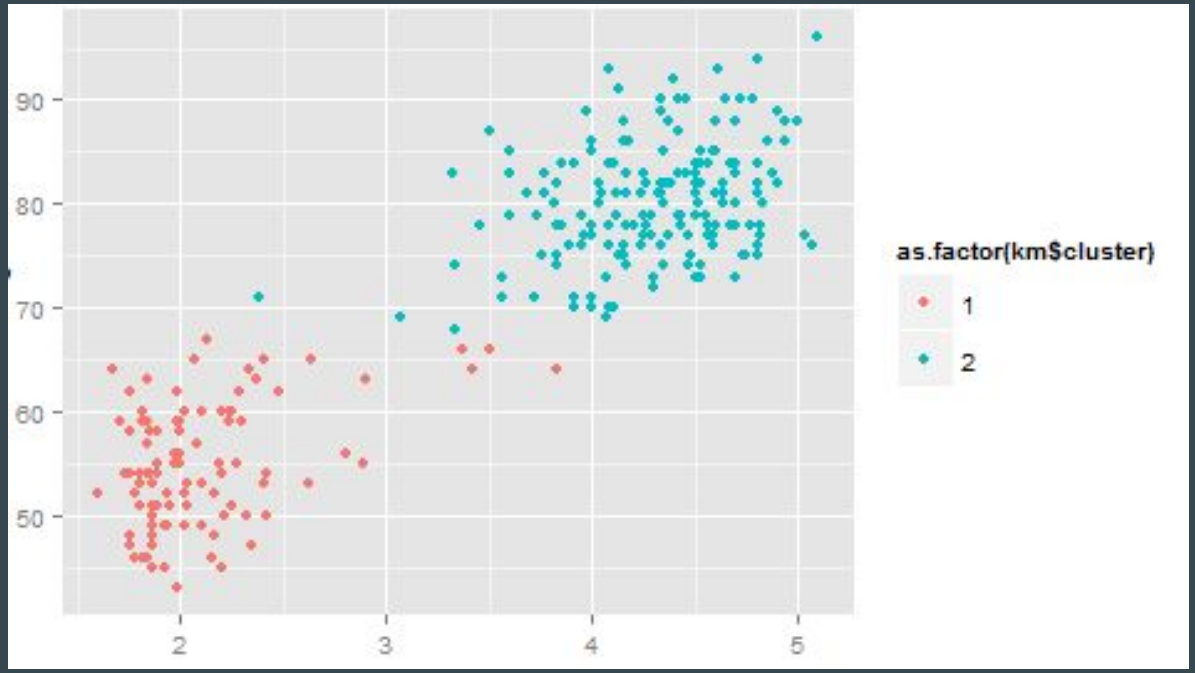
How do we deal with spatial dependence?

- Time is ordered: t=0 comes before t=1; t=2 comes after, and so on
- Space is not ordered
  - Florida does not come before Arkansas
- Remember, the problem with spatial dependence is that, uncorrected, it will lead to us over estimating how much information we have
- Instead we calculate the standard errors for each cluster

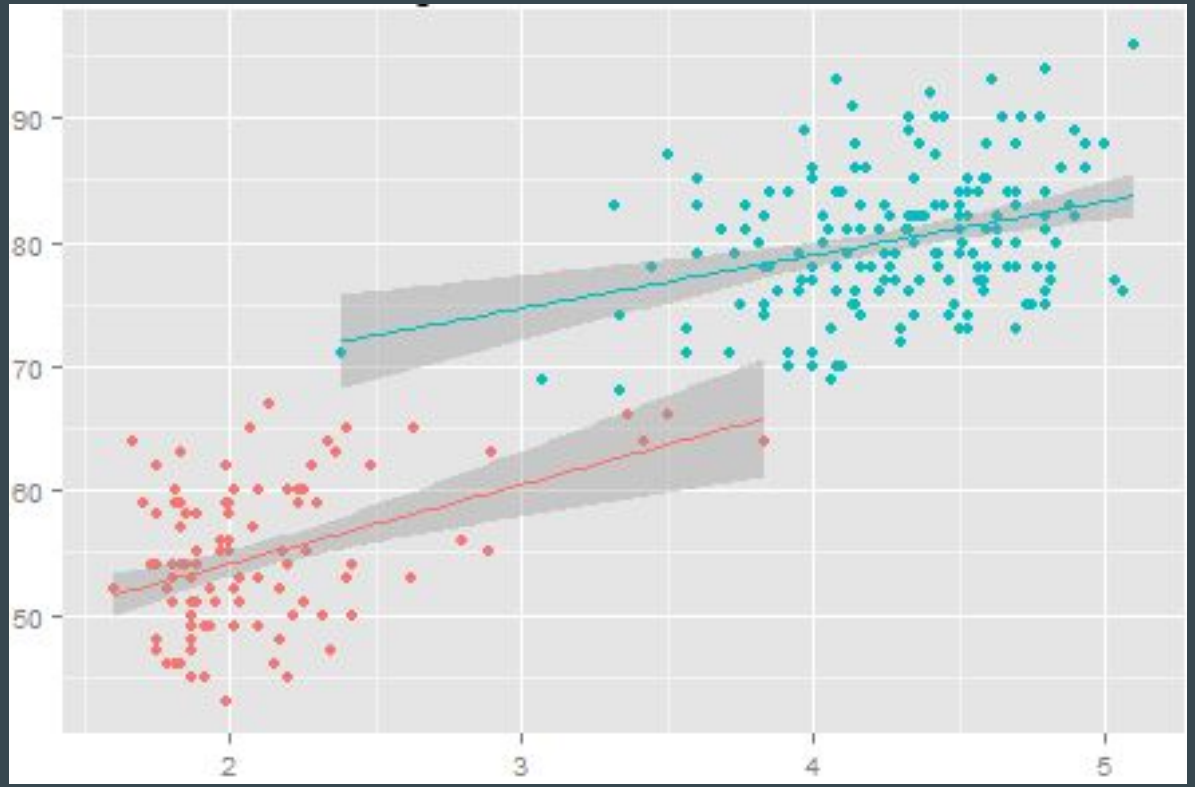
Clustered standard errors

- When we compare across spatial clusters, we over estimate the amount of information we have
- However, when comparing observations within cluster, we hold the spatial cluster constant
- What if we just estimated our standard errors for one cluster at a time, and then combined them at the end?

Clustered standard errors in action



Clustered standard errors in action



Summary

- We learned that the main assumption that we make for linear regression is additivity / independence
- These assumptions can be violated in multiple ways
- Social data is full of complicated dependence structures

Summary

Violation of independence can come from many sources

1. Interactions between our explanatory variables
2. Temporal dependence in our outcome variable
3. Spatial dependence in our outcome variable

Summary

- However, there are ways to incorporate these dependence structures in our models
- We can use interaction terms between variables that have a conditional effect on one another
- We can use lagged outcome variables to adjust for temporal dependence
- We can use clustered standard errors to account for spatial dependence