Methodological Notes

Age, Cohort and Period Effects

Age Effect (life-cycle effect)

- Age effects are changes that are the result of growing older
  - E.g. Partisan attachment as one grows older (Tilley 2003)

Cohort Effect (generational effect)

- Age effects deal with changes throughout people's lives but cohort effects deal with the stability of attitudes and behaviours over the course of people's lives
- "People acquire certain values and behavioural habits at early stages in their lives, and as a result of these socialisation processes some of these values and behavioural habits tend to be very stable" (van der Brug, Hobolt and de Vreese 2015)
- If these attitudes and behaviours are indeed very stable, aggregate changes would be mainly the result of generational replacement.
- Generational effect: Differences between age cohorts that persist over a long period of time
  - E.g. the older generation likes to chew betel leaves but the younger generation do not; over time, the number of people who chew betel leaves decrease due to a generational/cohort effect

Period Effect

- Period effects are changes in behaviours and attitudes that affect all generations
  - E.g. a ban on whaling makes less people consume whales

Related Issues

- Socialisation studies often struggle to disentangle the three effects statistically, because differences over time (period effect), differences between generations (cohort effect) and differences between age groups (age effect) are largely confounded
- Solution: combine the following two strategies
  - (1) Define testable hypotheses that are theoretically grounded
  - (2) Evaluate the empirically observed patterns from each of the three perspectives
- The literature suggests that changes result mainly from a combination of period effects and generational replacement for most politically relevant attitudes and behaviours (van der Brug, Hobolt and de Vreese 2015)

Co-efficients Interpretation

- Source: https://www3.nd.edu/~wevans1/econ30331/interpreting_coefficients.pdf

Top Mistakes with Regression Analysis

- Source: https://sites.google.com/site/modernprogramevaluation/variance-and-bias
- An estimator is **consistent** if, as the sample size increases, the estimates (produced by the estimator) "converge" to the true value of the parameter being estimated.
  - Consistency means that, as the sample size increases, the sampling distribution of the estimator becomes increasingly concentrated at the true parameter value.
- An estimator is **unbiased** if, on average, it hits the true parameter value. That is, the mean of the sampling distribution of the estimator is equal to the true parameter value.
  - Bias means that the expected value of the estimator is not equal to the population parameter
- **Endogeneity** occurs when a predictor variable (x) in a regression model is correlated with the error term (e) in the model
  - Common sources of endogeneity are omitted variable bias, simultaneity bias and measurement error

Multicolinearity

- Occurs when the independent variables in a regression model are very strongly correlated with each other
- This makes it difficult to tell the independent effects of those variables
- Does not create bias but gives unstable coefficient estimates (adding and removing data changes the estimates a lot; e.g. using data person 1-200 vs. 1-199 and person 201 gives very different estimates)
- Standard errors of each slope are inflated
  - When the standard errors are larger the confidence intervals are bigger (lower precision) and it is less likely that the slope will be statistically significant
- High R^2 but will have insignificant coefficients
- Perfect multicollinearity: occurs when two or more independent variables have an exact linear relationship between them
  - One value can be predicted from the other values
  - Perfect multicollinearity can arise as a result of the Dummy Variable Trap where a redundant dummy variable is added (e.g. using dummy variables for both Male and Female)
- We can detect if variables are collinear by finding pairwise correlations among IV; if correlation coefficient > 0.8 then severe multicollinearity may be present
- Solution
  - If multicollinearity arises due to measurement, it can be solved by omitting a variable
    - E.g. omitting one of “weight in kg” or “weight in pounds”

- If it is causal, use better research design
  - Respecify the model by combining multi collinear variables (e.g. combine GDP and Population and instead use GDP per capita)

Omitted Variable Bias

- Occurs when a variable that is correlated with both the dependent variable and one of the included independent variables is omitted from the regression
  - If the omitted variable is not correlated with another independent variable at all, excluding it does not produce bias
  - The more the omitted variable is correlated with the IV, the larger the bias
- Example: estimating the effect of activity on bone density without including weight, but weight is negatively correlated with activity level; regression will produce a negatively biased coefficient
- Direction of bias

	Included and omitted IVs negatively correlated	**Included and omitted IVs positively correlated**
Included IV and DV negatively correlated	Positive bias: coefficient is overestimated	Negative bias: coefficient is underestimated
Included IV and DV positively correlated	Negative bias: coefficient is underestimated	Positive bias: coefficient is overestimated

Source: https://statisticsbyjim.com/regression/confounding-variables-bias/

- Omitting an explanatory variable from the regression model will increase the error variance, which suggests an increase in the variance of the OLS regression coefficients
- Omitted variable bias causes endogeneity because the effect the omitted variable has gets included in the error term but nonetheless is correlated to at least one of the IVs
- Solutions
  - Better research design (include all potentially relevant variables)
    - If the variable that should be included is hard to measure then consider using a proxy variable (a variable that is highly correlated with the actual variable you want to include e.g. GDP per capita as proxy for quality of life)
    - Omitted variable bias implies we have unknown differences in subjects; a way to control for differences within subjects is to randomly assign participants to a treatment or a control group (Randomised Control Trial; RCT)
    - BUT introducing the confounder to the regression causes multicollinearity since the existing and the added IVs are correlated; there is a trade-off between precision (from multicollinearity) and bias (from omitted variable)
  - Fixed effects
    - When RCT is not feasible, we can instead include fixed variables, usually by as dummy variables for the missing or unknown characteristics

- Fixed variables are constant or change at a constant rate over time across individuals
    - E.g. age, sex, or ethnicity; any effects from being a woman, a person of colour, or a 17-year-old will not change over time, or at least change at a very slow rate
- Fixed effects models remove omitted variable bias by measuring changes within groups across time

- Instrumental variables
    - The omitted variable bias arises because the variation in X is not independent of the error
    - We can introduce an instrumental variable that splits the explanatory variable X into two parts – the part that could be correlated with the error term and the part that probably is not (Z)
    - An instrumental variable (Z) is uncorrelated with the error term (e) (i.e. Z is exogenous) but is correlated with X (relevance); Z is correlated with Y but only indirectly through X (excludability)
        - E.g. proximity to college (Z) might be correlated with schooling (X) but not with wage residuals (e); wage (Y)
        - Common sources of instrumental variables:
            - Nature (e.g. geography, biology, weather) – a truly random source of variation that influences X
            - History – things determined a long time ago which no longer plausibly influence Y but still influence X
        - Note: While proxy variables appear in the structural equation, Z does not (due to the condition of exclusion)
    - We can find the coefficients on X and Z using indirect least squares or two stage least squares (2SLS)
    - The estimator on the instrumental variable Z captures only the effects of shifts in X induced by Z on Y whereas the OLS estimator captures not only the direct effect of X on Y but also the effect of the omitted variable
        - Having information on the coefficient of X and the coefficient of Z is sufficient for calculations of the exogenous effect of X on Y
    - Since variation of Z is independent of the DV by exogeneity and any correlation with X has already been accounted for, the omitted variable bias problem is solved

DV Measurement Error

- Measurement Error
    - Variables in social science are often very hard to measure accurately, sometimes due to
        - Practical constraints in obtaining data (e.g. number killed in a civil war)
        - Conceptual imprecision (e.g. democracy scales)

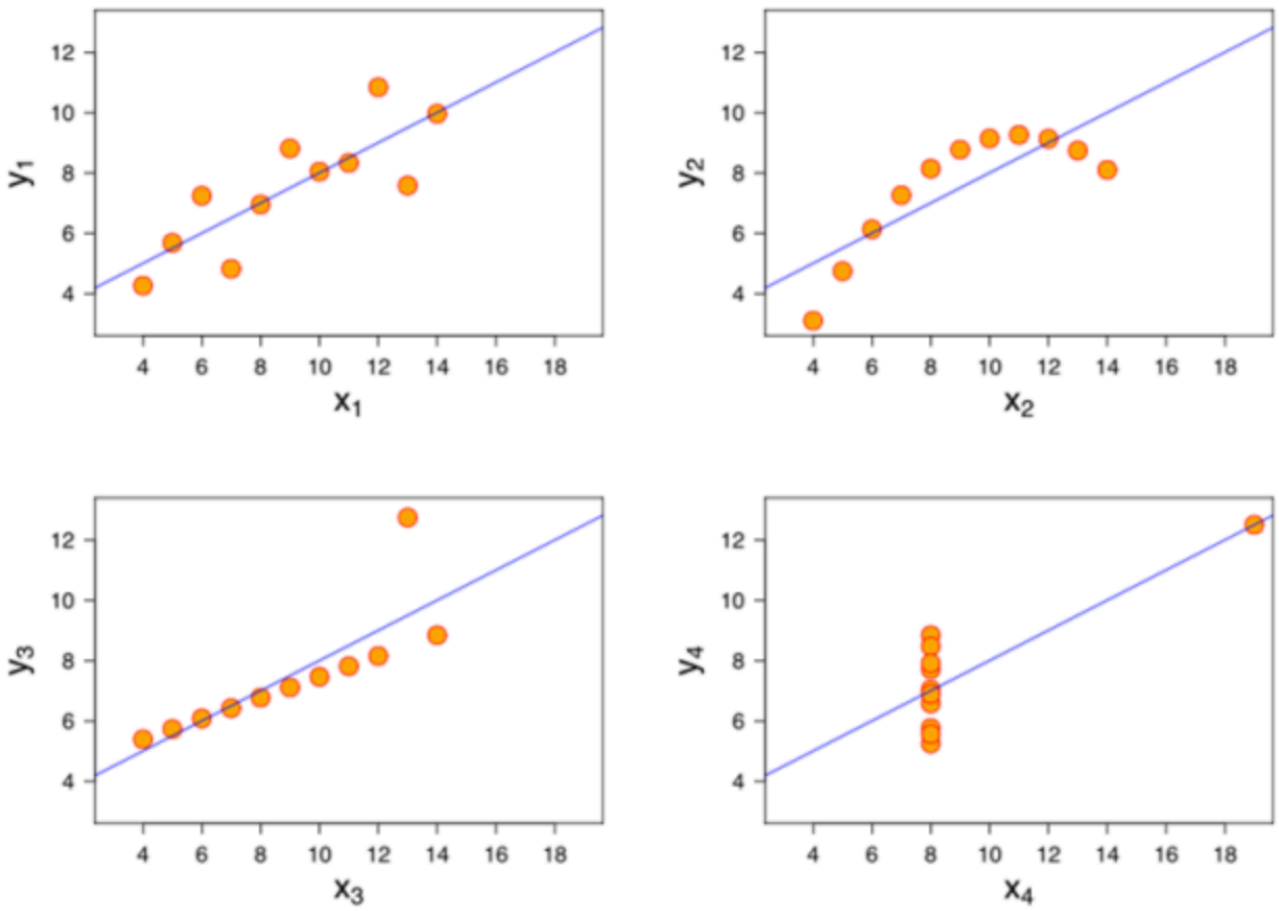
    - A measurement error is a random error not a systematic error
        - A systematic error would shift the mean and the intercept but the coefficients will not be biased
    - Data transformations tend to magnify the measurement error problem
- A DV measurement error will cause the dependent variable (Y) to have greater variance, so the standard error will be bigger
- DV measurement error does not produce a bias; slope coefficients will be identical
- However, the inflated s.e. will make it less likely that the slope will be statistically significant
- Solutions
    - Use better measurement
    - Increase sample size (higher sample size tends to reduce the variance of the variable we are measuring)

IV Measurement Error

- An IV measurement error will cause the independent variable (X) to have greater variance
- This shrinks the slope estimate towards zero (attenuation)
- The standard errors unpredictable but incorrect
- Solutions
    - Use better measurement
    - Increase sample size (higher sample size tends to reduce the variance of the variable we are measuring)

Misspecification Bias

- Using an inappropriate form of the proper regression model for the variables under analysis can introduce bias
- Misspecification can affect the standard error and cause the coefficient to be biased
- Anscombe's quartet



- Solution
  - Visualise the data to see which regression model best matches the data
  - Modify the data with non-linear terms such as squared and cubed terms or log transformations

Outliers

- Outliers are data points that fall far away from the major “cluster” of points
  - They can be actual data points or erroneous values
- Outliers may make the line of best fit deviate from the major cluster of points, thereby causing the coefficient estimates to be biased
- If the coefficients are biased, the standard errors will also deviate from the true value
- Solution
  - Visualise the data
    - View points on a graph
    - Use a box plot to see points that are far away from the quartiles
    - Find isolated bars on histograms
  - Sort data to identify outliers

Heterogeneity Bias (Group Difference Bias)

- Heterogeneity bias occurs when there are natural group structures in the data and there are innate differences in the groups that are correlated with the dependent and independent variables
- Heterogeneity bias is a special case of the omitted variable bias where the omitted variable is groups
- As a kind of an omitted variable bias, heterogeneity can cause the estimated coefficient to be biased as well as affect the standard error
- Solution
  - If the observed differences are due to unobservable variables (e.g. intelligence, natural ability, efficiency in task completion), use fixed effects

Selection Bias

- Selection bias arises when individuals enter/exit groups in non-random ways
  - E.g. sending survey about product to newspaper subscribers who may be different from paying audience
  - Self-selection bias: occurs when subject of studies select themselves
    - E.g. estimating the average wage of women using data collected from a population of women, but housewives were excluded by self-selection
- Selection bias is a special case of the omitted variable bias where the omitted variable is propensity to participate
- Lower internal validity: estimated coefficients can be biased in either direction
- Lower external validity: selection bias can mean the sample is unrepresentative
- Larger standard error
- Solutions
  - Control non-random selection through research design (e.g. do RCT with treatment and control groups)
  - Model the selection process using econometric methods such as the Heckman Selection Model

Random Attrition

- Attrition bias is a kind of selection bias, which is a kind of omitted variable bias
- Attrition bias occurs when participants drop out of the study (e.g. participants in a medical trial not turning up for follow ups)
- The remaining group can become unrepresentative
  - Selective attrition bias occurs when there are differences between treatment and control groups (as supposed to differences within the same group from the beginning and towards the end of the study)

- Concerns arise when attrition rates are over 20%; Schulz and Grimes (2002): attrition rates under 5% is usually of no concern
  - BUT a study with low attrition rate might be more susceptible to bias than a study with a higher attrition rate if the drop-outs have very unique characteristics
- However, if attrition is random (uncorrelated) then it is not a problem
- Random attrition does not introduce bias but can increase the standard errors (since $s.e. = \frac{\sigma}{\sqrt{N}}$ and attrition causes N to decrease)
- Solutions
  - Experiment design
    - Keep follow-up interviews as brief as possible
    - Offer incentives (e.g. cash, gift cards)
    - Use a good tracking system with detailed contact information
    - Remind participants of appointments with postcards and telephone calls

Simultaneity Bias

- Simultaneity occurs when two variables on either side of the regression influence each other simultaneously; X causes Y but Y also causes X
  - This is distinct from the problem of reverse causality where X is said to cause Y but in fact Y causes X
- The causal structure forms a feedback loop making it very difficult to separate out the independent effects
  - E.g. Criminals choose crime scene based on number of police around, but police also choose where to station based on the crime rates
- Simultaneity can cause the estimated coefficients to be biased and also affect the standard error
- Solutions
  - Use lagged variables if panel data is available
  - Structural Equation Models (SEM) (i.e. simultaneous equation models)
  - Instrumental variables