Regression – Selected topics

Research Master Course

Lecture 2

September 9 2011

AGENDA

- Regression – selected topics:
  - Categorical and continuous X-variables
  - How does multiple regression control?
  - Causal models (indirect and confounding effects)
  - Interaction (moderator) analysis.
- Type I and type II errors.

Simple regression

- Y = a + b.X
- Y = B0 + B1*X1
- The model informs how the expectation (mean) of Y is related to X1.
- OLS estimation: the estimates of B0 and B1 minimizes Sum(Yi-^Y)**2. It is unimportant to know how this works.
- OLS gives relatively much weight to outliers.

Intercept

- Intercept = where the regression line intersects the Y-axis (where X=0).
- Intercepts are often totally unimportant for sociological interpretation; however…
- Make sure that you always understand why the intercept is as low / high as it is.
- If X does not include 0 as a value, B0 is an ***extrapolation*** and be meaningless. To make a model better interpretable it often helps to transform X to include 0:
  - As the minimum value
  - As the mean / middle value (centering).

Slope

- The slope B1 informs how much Y you gain/lose for 1 unit of X.
- It helps to make the unit of X interpretable:
  - Transform to 0..1 or 0,1: slope value covers the entire range X.
  - Divide large values (age, years, income) by a constant to see significant digits.
  - Standardize X to Z-scores, if the unit of X is arbitrary to begin with.

Regression with dummy variables

- We use dummy variables to represent a nominal X-variable.
- If X has K categories, K-1 dummy effects can be estimated. The omitted dummy effect constitutes the reference category.
- The estimated dummy effects are the differences from the reference categories.

Reference category (1)

- Regression programs would normally drop one dummy effect. Exception: pairwise deletion of missing values.
- It may be wise to make you own choice.
- Possible choices:
  - Largest category
  - “Middle” category
  - First category
  - Last category.
- It does not make a real difference what you choose, but make sure:
  - That you know which category is omitted
  - That this category is not extremely small.

Reference category (2)

- It is important that not only you, but also your reader knows the reference category.
- Unfortunately, there is no standard way of doing this. Alternatives:
  - Footnote to table
  - ***Listing the effect as zero***
  - Listing the effect as “reference”.
- *Never omit the reference to the omitted category.*

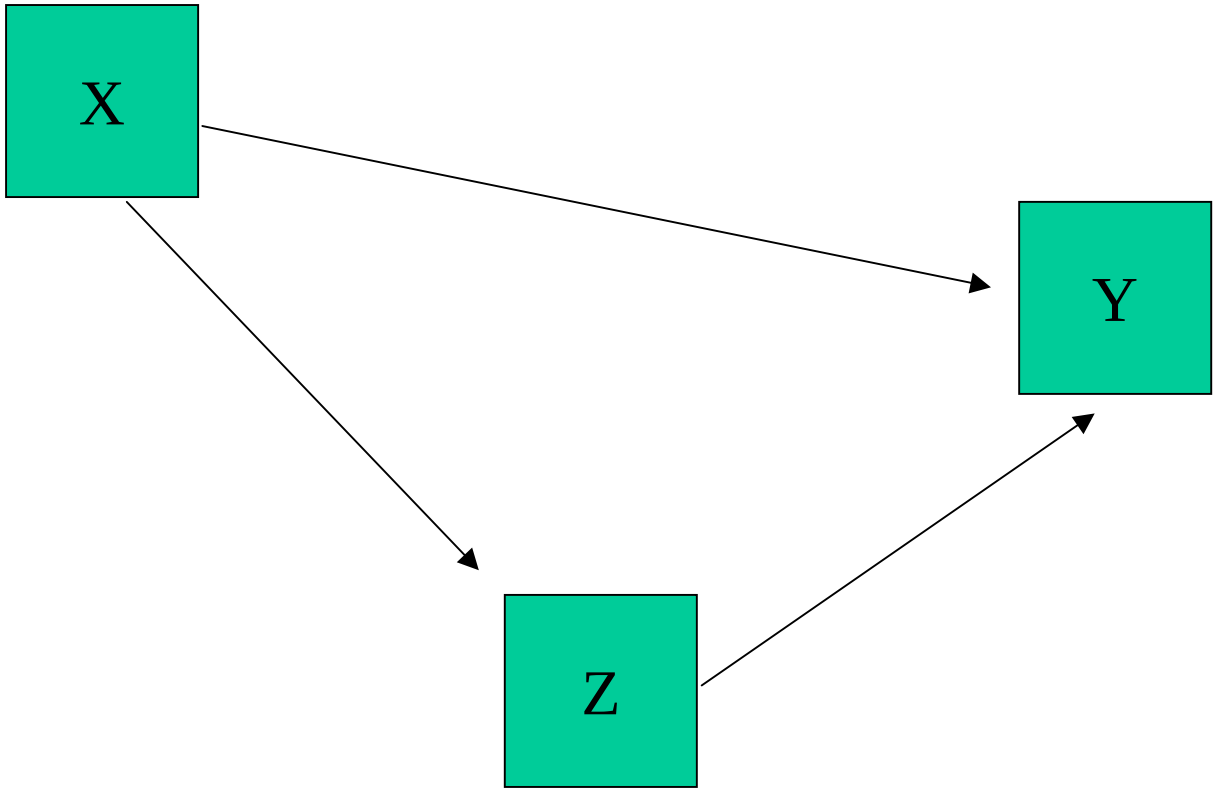
Multiple regression

- Y = B0 + B1*X1 + B2*X2.
- B1 estimates the effect of X1, controlling (keeping constant) the effect of X2; B2 estimates the effect of X2, controlling (keeping constant) the effect of X1.
- B1 en B2 are partial effects.
- But how does it do this?

How does multiple regression control?

- Step 1: Regress Y on X1 and X2.
- Step 2a: Regress X2 on X1. Compute residuals cX2 (=take out effect of X1).
- Step 2b: Regress X1 on X2. Compute residuals cX1 (=take out effect of X2).
- Partial effects of step 1 are equal to simple effects of cX1 and cX2 on Y.

The elementary causal model (1)



The elementary causal model (2)

- Variables:
  - Y: dependent variable
  - X: confounder for Z→Y
  - Z: mediator for X→Y
- Total correlation =

  ***direct effect* + *indirect effect* + *confounding effect*.**
  - r(XZ) = b(X→Z)
  - r(XY) = b(X→Y) + b(X→Z)*b(Z→Y)
  - r(YZ) = b(Z→Y) + b(X→Z)*b(X→Y)

Confounding and mediation: causal order

- Whether variables are confounders or mediators, cannot be decided with statistics.
- Causal order is an assumption that must be argued from the research design:
  - Timing arguments (e.g. history, life cycle, retrospective questions, panel design).
  - General variables (attitudes) influence specific variables (opinions, behavioral choices), not the other way around.

Interaction (1)

- In interaction [moderator] models, we investigate whether the effect of X1 depends upon the value of X2 (vice versa).
- Y = B0 + B1*X1 + B2*X2 + B3*X1*X2
- For easy interpretation of the model is very convenient when both X1 and X2 contain an interpretable value 0 and unit 1:
  - Score between 0..1, or 0,1
  - Z-score: M=0, SD=1.

Interaction (2)

- X1=0, X2=0: Y = B0
- X1=1, X2=0: Y = B0 + B1
- X1=0, X2=1: Y = B0 + B2
- X1=1 : Y = (B0 + B1) + (B2+B3)*X2
- X2=1 : Y = (B0 + B2) + (B1+B3)*X1
- B1: effect of X1 if X2=0
- B2: effect of X2 if X1=0
- B3: how effect of X1 changes if X2 increases by 1
- B3: how effect of X2 changes if X1 increases by 1.

Pitfalls

- Occasionally interaction models do not converge if units of variables are very different.
- Interaction (as above) is still about the effects of two variables. Despite the presence of three terms, it is only two variables. It is nonsensical to interpret the terms independent of one another.
- Never leave out the main effects of X1 and X2. It makes the model uninterpretable.

Statistical Significance and Statistical Power

Type 1 and Type 2 errors

- H0: Assumed exact population value, usually H0=0.
- H1: usually unspecified alternative, such as:
  - There is an effect
  - There is a positive effect There is a positive effect of .30 of stronger.
- Usually, the researcher's sympathy are with H1 – researchers would like to see the H0 rejected.

Two types of errors

- If H0 is true: Reject H0 – type I error.
- If H1 is true: Not reject H0 – type II error.
- We *choose* the probability (risk) of type I error. It is called the ***significance*** level (alfa). 5% is the standard choice.
- In the long run, we commit type I errors in 5% of all decisions: we know exactly *how often* we make this type of error, but *NOT when* we make it.
- 5% is a somewhat arbitrary choice (originally made by RA Fisher), there is no reason why it could not have been something else.

Type II errors

- The probability of making a type II error is called beta. 1-beta is called ***statistical power***.
- Beta, and (1-beta) are probabilities, ranging between 0 and 1.
- Generally we do not know how large beta is, but we do now circumstances in which beta is larger or smaller.
- To learn about some principles, it is useful to study the power graph, that you find in any good statistics book.

Increasing statistical power

- Larger N.
- Higher (!) alfa.
- One-tailed test (=directed H1).
- More extreme H1.
- More explained variance:
  - Matched samples, e.g. before after designs.
  - Enter important covariates, even if they do not correlate with X.
  - Constrained estimation.

Calculating statistical power

- We cannot calculate beta, if we do not have an exact H1.
- One approach is to set an arbitrary value of the statistic of interest.
- This is most easily done in a standardized statistic, such as the correlation coefficient.
- Example: what is the required sample size that you would need for a correct decision on r=.25 (‘moderate effect’) with 80% statistical power?
- Power calculations are most often stated in terms of minimum sample size. This is important, but remember there are other ways to increase power!!

One-degree of freedom test

- Statistical power is also influenced by the complexity of the model you estimate:
  - Degrees of freedom / number of parameters estimated
  - Complications by interactions and collinearity.
- By simplifying models to parsimoneous models, we do not only enhance interpretation – but also statistical power.
- Of course there is a trade-off between fit of the model the data and parsomony.