E IS FOR ENDOGENEITY

Endogeneity in Content Analysis & Finding Strong Instruments to Resolve It

John R. Busenbark

Academy of Management
2017 Content Analysis PDW
August 5th, 2017
Content analysis *quantifies text* to *measure* psychological *constructs* – what people think, feel, and do – by what they say

- **Many empirical benefits**
  - Psychometric advantages (Aguinis & Edwards, 2014)
  - Internal validity compared to archival (Short et al., 2010)
  - External validity compared to self-reports (McKenny et al., 2012)
  - More proximal than archival (McKenny et al., 2016)

- **Detriments owing to endogeneity**
  - Omitted variable bias (Semadeni et al., 2014)
    - Are subjects thinking something else that relates to the variable?
  - Measurement error (Kennedy, 2008)
    - Does the measure remain the same across many contexts?
Endogeneity in Content Analysis

- Relationship between regulatory focus and M&A activity

\[ y_{\text{(M&A activity)}} = \alpha + \beta_1(\text{Reg.Focus}) + \beta_2(\text{n(controls)}) + e \]

Everything correlated with \( y \) (net income) not included in the model

Maybe regulatory focus is really M&A confidence

\[ \text{Corr}[x,e] \neq 0 \]

“Instead of estimating the ‘true’ relationship between the independent and the dependent variable, OLS regression \underline{mistakenly includes} the correlation between the independent variable and the error term” (Semadeni et al., 2014: 1072)
The Solution: Two-Stage Models

First Stage

\[ x_{(\text{Reg.Focus})} = \alpha + \beta_{1-2(\text{instruments})} + \beta_{3-n(\text{predictors})} + u \]

Second Stage

\[ y_{(\text{M&A})} = \alpha + \beta_{1(\text{predicted x})} + \beta_{2-n(\text{controls})} + e \]

- **It ALL depends on instruments**
  - Corrected model must be specified correctly!
    - “Weak instruments can report results that are inferior to those reported by OLS…” (Semadeni et al., 2014: 1070)
    - No instruments “can often do more harm than good” (Kennedy, 2008: 271)

- **Good instruments are hard to find these days….**
  - MUST be exogeneous (Kennedy, 2008)
  - MUST be strong predictors (Stock et al., 2002)
  - Almost no SMJ articles from 2005-2012 had good ones (Semadeni et al., 2014)
    - “Completely inappropriate” tests for good instruments (Larcker & Rusticus, 2010: 192)
• **Instruments in content analysis**
  - Several content-analysis related instruments
  - Finding the right one depends on the context
    - Whether the content analysis constructs is the IV, DV, or both

• **Independent variable is from content analysis**
  - E.g., **Regulatory focus** predicts M&A activity (Gamache et al., 2015)
  - E.g., **Mentioning phrases** predicts analyst downgrades (Busenbark et al. forthcoming)
    - Number of characters in a document
    - Number of documents released
    - Number of relevant sections in the document
    - Size of the document/size of images in the document
    - Other content analysis constructs discriminant from the DV
Dependent variable is from content analysis
- E.g., Firm wrongdoing on tenor of media coverage (Zavyalova et al., 2012)
- E.g., CSR and media praise (Petrenko et al., 2016)
  - Archival financial data
  - Arellano-Bond estimation
  - Industry characteristics
  - Geographic characteristics

Independent and dependent variables from content analysis
- E.g., Positive media predicts more positive media (Pollock et al., 2008)
- E.g., Entrepreneurial language predicts legitimacy (Gao et al., 2016)
  - Number of characters in IV documents
  - Financial data theorized to predict IV
  - Arellano-Bond estimation if outcome is to an event
  - Executive demographic characteristics
A Couple Caveats on these Instruments

- Some can represent important constructs
  - Number of characteristics reflects info complexity (Loughran & McDonald, 2014)
  - Number of documents can reflect litigation or abnormal behavior (Donelson et al., 2012)
  - Ensure industry or geography is not an important component of the DV construct (Wang et al., 2014)
  - Arellano-Bond estimation is requires great precision (Arellano & Bond, 1991)

- Don’t jump into two-stage models too quickly
  - They’re not the most efficient
  - May present significant Type II errors
  - Think through your content analyses constructs conceptually
Thank You!

ANY QUESTIONS?


