What is Wrong With “Hypotheses Sociology”? Or: How Theory-Driven Empirical Research Should Look Like

Katrin Auspurg and Josef Brüderl

November 2016
Social Research in the “Era of Regression”

• Since the advent of regression, social researchers struggle with how to best use these statistical tools

• In the 1970ies many social researchers used regression “Y-centered”: they threw in many variables to “explain” variance

• This a-theoretical practice was criticized by many. Instead it was suggested to guide variable selection by theory
  – Theory-driven empirical research
  – However, the practical implementation of theory-driven research often looked like this: researchers used one/several theories, deducted several hypotheses, and simply put all variables in the regression (“hypotheses sociology”)

• Some authors argue that hypotheses sociology is often misguided
  – G. King (1986) How Not to Lie with Statistics
  – F. Elwert (2016) Comments On Backdoor-Based Identification
Fundamental Rules of Causal Inference

• Our research problem
  – Identifying a causal effect

• Control for confounders
  – If you do not, you have an omitted variable bias
  – If some confounders are unobserved one has to use methods like IV, FE or RD

• Do not control for colliders
  – If you do, you have an endogenous selection bias

• Do not control for mediators
  – If you do, you have an overcontrol bias
  – [If you want to get at the total causal effect]
Hypotheses Sociology

• We are interested in the determinants of some outcome Y
• We use one/several theories to derive hypotheses
  – H1: D affects Y positively
  – H2: A affects Y negatively
• Then we estimate the following regression
  \[ Y = \alpha + \beta D + \gamma A \]
  – \( \beta \) is the causal effect of D ("controlling for A", or "net of A")
  – \( \gamma \) is the causal effect of A ("controlling for D", or "net of D")
• The fundamental problem of this strategy
  – It works only if the causal structure is of the type “multi-causality”
Hypotheses Sociology

- It no longer works if the causal structure deviates from multi-causality. For instance:

  \[ Y = \alpha + \beta D + \gamma A \]

  - Here, only \( \beta \) is a (total) causal effect
  - \( \gamma \) is only the direct effect, left after controlling for the mediator \( D \)
  - Thus, it would be erroneous to interpret \( \gamma \) as a total causal effect
    - Nevertheless, this erroneous interpretation is applied by many users

  - Obviously, this is a dramatic insight as much regression based empirical results are likely to be misinterpreted!
Regression needs a Causal Structure

• Here is another example (adapted from Elwert, 2016)

\[
\ln(\text{Wage}) = \alpha + \beta \text{Exp} + \gamma \text{Educ} + \delta \text{Female}
\]

• \(\beta\) is a total causal effect (all non-causal paths are blocked)
• \(\delta\) is the direct causal effect (mediator Exp controlled)
• \(\gamma\) is the direct causal effect (mediator Exp controlled) that is confounded (by unobservable U)
• Thus, it would be misleading to interpret each coefficient as a total causal effect
Regression needs a Causal Structure

- For identifying **one** causal effect we need **one** specially tailored regression model
- To estimate the causal effect of “Exp”
  \[\ln(Wage) = \alpha + \beta \text{Exp} + \gamma \text{Educ} + \delta \text{Female}\]
- To estimate the causal effect of “Female”
  \[\ln(Wage) = \alpha + \delta \text{Female}\]
- To estimate the causal effect of “Educ”
  \[\ln(Wage) = \alpha + \gamma \text{Educ} + \delta \text{Female} + U\]
  - Somehow one would have to account for the unobservable U
Controls

• Often one adds “controls”
  
  \[ Y = \alpha + \beta D + \gamma A + \delta Z \]

• Certainly, the effects of controls should not be interpreted as causal effects
  – \( \delta \) is not a causal effect! It is only the direct effect, left after controlling for the mediator D
  – Ironically, it works only if Z is not a confounder

• Nevertheless, this is often done
  – “Finally, let’s have a look at the effects of the controls …”

• N.B.: Often “confounders” are included without thought, e.g. “occupation”, “family type” (the usual suspects). Sometimes these are mediators, and will produce overcontrol bias
Current Social Research Practice

• Shortcomings of the standard “hypotheses-driven” social research article:
  – Theory is used to derive hypotheses on the effects of a number of variables on the outcome. But mostly nothing is said on the (complete) causal structure
  – Thus theorizing is only “loosely” coupled to the research problem
  – “Controls” are entered usually without theoretical arguments
  – Therefore, it is highly likely that some of the fundamental rules are violated and that estimates will be biased / misinterpreted
Lessons

• Don’t trust any article that infers many effects from a single regression without theorizing the complete causal structure of the research problem
  – Start yourself thinking about the causal structure. Draw a DAG.
  – From that you might be able to infer which effects are identified

➔ Don’t trust most regression based social science articles

• Stop teaching the hypotheses-driven approach to social research
  – Start teaching a “new style to causal analysis”
The New Style of Causal Analysis

- Focus on just one causal effect (X-centered)
  - What is the causal effect that your research problem aims at?
- Theorize on the complete causal structure
  - What are confounders, what are colliders?
  - Draw a DAG representing the causal structure
- Theorize on the intervening mechanisms (mediators)
  - No causation without a plausible mechanism
  - In the first step do not control for mediators (overcontrol bias)
  - Use them in a second step to explain the causal effect
- Think about identification
  - Given the causal structure, how can I identify the causal effect?
An Example for a Hypotheses-Driven Paper

• Authors BPZ investigate the factors that affect the survival chances of newly founded business firms (published in ASR)
  – Outcome: business failure rate

• Theories used to derive hypotheses
  – Human capital theory
  – Organizational ecology

• Hypotheses:
  – “We expect more schooling to improve a firm's survival chances”
  – “We expect work experience to show a decreasing payoff”
  – “Size at time of founding should increase survival chances”
  – …

  – Altogether 19 hypotheses (“a rich set of hypotheses”)!
An Example for a Hypotheses-Driven Paper

- The authors present one regression
- They interpret each coefficient as if it is a (total) causal effect

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Coefficient</th>
<th>t-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Capital</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of schooling</td>
<td>-.054</td>
<td>3.18</td>
</tr>
<tr>
<td>Years of work experience</td>
<td>-.051</td>
<td>3.92</td>
</tr>
<tr>
<td>Years of work experience squared/100</td>
<td>.101</td>
<td>2.97</td>
</tr>
<tr>
<td>Industry-specific experience</td>
<td>-.332</td>
<td>3.53</td>
</tr>
<tr>
<td>Self-employment experience</td>
<td>.096</td>
<td>.86</td>
</tr>
<tr>
<td>Leadership experience</td>
<td>.190</td>
<td>1.39</td>
</tr>
<tr>
<td>Self-employed father</td>
<td>-.105</td>
<td>1.09</td>
</tr>
<tr>
<td><strong>Organizational Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follower business</td>
<td>-.446</td>
<td>3.46</td>
</tr>
<tr>
<td>Affiliated business</td>
<td>.278</td>
<td>2.06</td>
</tr>
<tr>
<td>Amount of capital invested</td>
<td>-.034</td>
<td>3.40</td>
</tr>
<tr>
<td>Number of employees natural log</td>
<td>-.451</td>
<td>5.01</td>
</tr>
<tr>
<td>Registered in commercial register</td>
<td>-.793</td>
<td>4.48</td>
</tr>
<tr>
<td>Specialist business</td>
<td>-.189</td>
<td>1.94</td>
</tr>
<tr>
<td>Innovative business</td>
<td>-.133</td>
<td>1.19</td>
</tr>
<tr>
<td>National market-scope</td>
<td>-.363</td>
<td>3.59</td>
</tr>
<tr>
<td><strong>Environmental Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location in Munich</td>
<td>.131</td>
<td>1.49</td>
</tr>
<tr>
<td>In construction</td>
<td>.522</td>
<td>1.38</td>
</tr>
<tr>
<td>In wholesale/retail trade</td>
<td>.512&quot;</td>
<td>2.74</td>
</tr>
<tr>
<td>In transportation</td>
<td>.765&quot;</td>
<td>3.19</td>
</tr>
<tr>
<td>Restaurant business</td>
<td>.227</td>
<td>.87</td>
</tr>
<tr>
<td>In computer services</td>
<td>.486</td>
<td>1.81</td>
</tr>
<tr>
<td>In other services</td>
<td>.249</td>
<td>1.32</td>
</tr>
<tr>
<td>Competition intensity</td>
<td>-.205</td>
<td>1.39</td>
</tr>
<tr>
<td>Seasonality</td>
<td>.132</td>
<td>1.45</td>
</tr>
<tr>
<td>Clustering of orders</td>
<td>-.519&quot;</td>
<td>3.90</td>
</tr>
</tbody>
</table>
An Example for a Hypotheses-Driven Paper

• Some theoretical thoughts on the causal structure of the research problem show that the structure very likely is not of the „multi-causality“ type

![Diagram]

- Work Experience → Firm Size
- Schooling → Work Experience
- Schooling → Failure
- Firm Size → Failure
- Firm Size → Schooling

• Given this causal structure, the regression presented by the authors is plagued by an overcontrol-bias concerning the effect of “schooling”