CPS Linking & Validation
Linking CPS Data

Opportunities and Challenges
Opportunities

• Large sample sizes
• Good coverage of subpopulations
• Short-run changes in employment and families; reactions to births and losses (death, divorce)
• Combine rich sources of information about different topics
  • Work and volunteering; veterans and employment
Same Month One Year Apart

• Useful for studying year-to-year change
  • Earnings and employment dynamics,
  • Geographic mobility,
  • Movement into and out of labor unions

• Researchers can expect:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Linked N</td>
<td>54,890</td>
<td>48,131</td>
<td>53,486</td>
</tr>
<tr>
<td>Retention Rate</td>
<td>74.1%</td>
<td>69.3%</td>
<td>78.8%</td>
</tr>
</tbody>
</table>
Single Cohort, All 8 Months

• Useful for studying short-term dynamics
  • Change in economic arrangements as a function of social and demographic characteristics
  • Changes in these relationships over time

• Researchers can expect:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Linked N</td>
<td>11,012</td>
<td>10,069</td>
<td>11,528</td>
</tr>
<tr>
<td>Retention Rate</td>
<td>59.0%</td>
<td>59.7%</td>
<td>68.0%</td>
</tr>
</tbody>
</table>
Challenges

• Rotation pattern
• Linking keys
• Same variables with different codes
• Non-response
• Another layer of complexity if ASEC is part of design
• Data management
Challenges: Rotation Pattern

- CPS is NOT a longitudinal survey that follows one or several cohorts as they age.
- CPS does have a panel component where individuals are observed up to 8 times over 16 months, but the way they move through the survey is in a 4-8-4 rotation.
Challenges: Rotation Pattern

• Enter IPUMS CPS RoPES

• Motivation
  • Easily see CPS rotation pattern
  • Explore what topical supplements can be examined together
Challenges

• Rotation pattern
• Linking keys
• Same variables with different codes
• Non-response
• Another layer of complexity if ASEC is part of design
• Data management
Challenges: Linking Keys

• Changes in variables needed to link
  • Technical changes
  • Privacy protection changes

• Duplicate and recycled identifiers
  • Feng 2001
  • Flood et al. 2020
Challenges: Linking Keys

<table>
<thead>
<tr>
<th>HRHID</th>
<th>HRHID</th>
<th>HRHID</th>
</tr>
</thead>
<tbody>
<tr>
<td>PULINENO</td>
<td>PULINENO</td>
<td>PULINENO</td>
</tr>
<tr>
<td>STATEFIPS</td>
<td>STATEFIPS</td>
<td>STATEFIPS</td>
</tr>
<tr>
<td>HUHHNUM</td>
<td>HUHHNUM</td>
<td>HUHHNUM</td>
</tr>
<tr>
<td>HRSAMPLE</td>
<td>HRSAMPLE</td>
<td>HRHID2</td>
</tr>
<tr>
<td>HRSERSUF</td>
<td>HRSERSUF</td>
<td></td>
</tr>
</tbody>
</table>

**Transformations detailed in Drew, Flood & Warren 2014**
Challenges: Linking Keys

• CPSID(P) – an IPUMS-created unique identifier
• Bridges changes in variable names as well as the technical aspects of how to link
• Available for 1976 forward
• Drew, Flood, & Warren, 2014 Journal of Economic and Social Measurement
• Flood et al., (forthcoming) Journal of Economic and Social Measurement
CPSID(P) Limitations

• Mechanical
  • Created solely based on linking keys
  • Does NOT condition on AGE, SEX, RACE

• Not available for all respondents in ASEC
  • only those who are also in the March BMS
CPSID(P) Limitations

- Unlinkable records
  - Records with duplicated identifiers in 1976-1983
  - Children are in some supplements but not basics in 1976-1981

- Unlinkable files
  - In 1976-1978, supplements and BMS-only months
  - Before and after July 1985
  - Before and after October 1985
  - Before and after June 1995
Future Linking Work

• Make ASEC oversamples linkable using CPSIDP

• A version of CPSIDP that is validated on AGE, SEX, RACE matches
  – Coming very soon!
Challenges

• Rotation pattern ✓
• Linking keys ✓
• Same variables with different codes
• Non-response
• Another layer of complexity if ASEC is part of design
• Data management
Challenges:
Same variables, different codes

- IPUMS!
- Harmonized across time to deal with changes in codes
- Use it for good, never for evil!
  - (Also, please cite us!)
Challenges

• Rotation pattern ✔
• Linking keys ✔
• Same variables with different codes ✔
• Non-response
• Another layer of complexity if ASEC is part of design
• Data management
Challenges: Non-response

• Households that were eligible to participate in CPS but did not because of
  • Non-response (at any point in the rotation)
  • Death
  • Migration

• Imputation
Challenges: Non-response

• COVID-19
  • Response rates dropped noticeably during the pandemic, especially for incoming rotation groups
  • Individuals who link across months during the pandemic are more likely to be White, non-Hispanic, and to have higher income than those who are eligible to link, but do not (Rothbaum & Bee, 2020; Ward and Edwards, 2020)
Challenges: Non-response

CPS BMS response rates by month, 2019-2021

Response Rate

Month
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

2019 total 2020 total 2021 total
Challenges

- Rotation pattern ✔
- Linking keys ✔
- Same variables with different codes ✔
- Non-response ✔
- Another layer of complexity if ASEC is part of design
- Data management
Challenges: ASEC

• Differently named variables, including linking keys, than basic monthly
• MARBASECID (1976, 1978+)
  • Variable to link ASEC and March basic monthly
  • Allows us to put CPSIDP on ASEC
• Flood & Pacas 2017 Journal of Economic and Social Measurement
• Flood et al., (forthcoming) Journal of Economic and Social Measurement
Challenges

- Rotation pattern
- Linking keys
- Same variables with different codes
- Non-response
- Another layer of complexity if ASEC is part of design
- Data management
Challenges: Data Management

• Large number of files
• Needing to enforce rotation pattern to guard against ID recycling

• Strategies for dealing with this include:
  • Use IPUMS!
  • Loops
  • Temporary files
  • Long data format
Challenges

- Rotation pattern
- Linking keys
- Same variables with different codes
- Non-response
- Another layer of complexity if ASEC is part of design
- Data management
Validation

Approaches, Our Assumptions, and a Note on Data Structure
Validation: Approaches

• Because CPSIDP is mechanical, this is a step we recommend you do if you are linking CPS data across months

• No right or wrong way to do this, but there are a couple approaches in the literature
  1. Use AGE, SEX, RACE (Madrian & Lefgren)
  2. Use Bayesian approach (Feng)
     • Argues that some people are lost with first approach; this yields higher linkage rates
Validation: AGE, SEX, RACE Rules

• SEX & RACE
  • No change allowed

• AGE
  • If both time points are in MIS 1-4 or MIS 5-8, difference between AGE at two time points is 0 or 1
  • If the first time point is in MIS 1-4 and the second is in MIS 5-8, difference between AGE at two time points is 0, 1, or 2
  • Beware of AGE topcodes!
    • codes 80 and 85! Allow for a greater AGE increase. Age 80=80-84. Age 85=85+. 
Validation: Age Rules

<table>
<thead>
<tr>
<th>AGE</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Can age 0 or 1 years compared to MIS 1
- Can age 0 or 5 years compared to MIS 1
- Can age 1 or 2 years compared to MIS 1
- Can age 0 or 5 years compared to MIS 1
Validation: Our Assumptions

• For however many observations you are linking
  • SEX and RACE may not change
  • AGE may only change in expected ways

• Changes in AGE are always compared to the first time point being linked
Validation: Our Assumptions

- AGE, SEX, and RACE must ALL match in expected ways – GOOD!

<table>
<thead>
<tr>
<th>MISH</th>
<th>AGE</th>
<th>SEX</th>
<th>RACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Six</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
<tr>
<td>Seven</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
<tr>
<td>Eight</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
</tbody>
</table>
Validation: Our Assumptions

- AGE, SEX, and RACE must ALL match in expected ways – BAD!

<table>
<thead>
<tr>
<th>MISH</th>
<th>AGE</th>
<th>SEX</th>
<th>RACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Six</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
<tr>
<td>Seven</td>
<td>21</td>
<td>Female</td>
<td>Asian only</td>
</tr>
<tr>
<td>Eight</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
</tbody>
</table>
Validation: For this Workshop

- Validation gets complicated quickly
- We want you to try to write validation code in the lab to think through validation
- We will provide validation code for you based on our all or nothing assumptions about AGE, SEX, and RACE for data in both
  - Long format
  - Wide format
Validation: Long Format

• For each individual
  • # records = # times observed

<table>
<thead>
<tr>
<th>MISH</th>
<th>AGE</th>
<th>SEX</th>
<th>RACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
</tbody>
</table>
Validation: Wide Format

• For each individual
  • # records = 1 regardless of # times observed

<table>
<thead>
<tr>
<th>AGE</th>
<th>SEX</th>
<th>RACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>Fem</td>
<td>White</td>
</tr>
<tr>
<td>21</td>
<td>Fem</td>
<td>White</td>
</tr>
<tr>
<td>21</td>
<td>Fem</td>
<td>White</td>
</tr>
</tbody>
</table>
### Long to Wide Format Data

**Long**

<table>
<thead>
<tr>
<th>MISH</th>
<th>AGE</th>
<th>SEX</th>
<th>RACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>Female</td>
<td>White</td>
</tr>
</tbody>
</table>

**Wide**
Questions?