Instrumental Variable Exercise

Effect of Centrally Acting ACEIs on incident dementia

- Paul L. Hebert, PhD
- Research Associated Professor
- Department of Health Services
- UW School of Public Health

Overview

- Creating a geographic-based instrument
- Testing the correlation with treatment
- Test the correlation with observed confounders
- Getting the IV estimate
  - Two Stage Least Squares (2SLS)
  - Two-stage Residual Inclusion (2SRI)
  - Bivariate probit (BP)

Creating instrument

- Geographic based instrument

  - Numerous studies have found unexplained small area variation in treatment patterns
  - Can these variations be used to create valid instruments?
    - Probably correlated with treatment
    - Are they uncorrelated with differences in potential outcome?
    - How do you define the “small geographic area” that is the basis of your instrument?
Creating instrument

- Approaches to defining the geographic area.

  - Use a predefined area
    The area resource file (ARF) maps counties to Health Service Areas (HSAs)
    Problem: Counties are huge and diverse
    Primary Care Service Area (PCSA)
    Dartmouth researchers have created aggregations of ZIP codes that match Medicare beneficiaries' travel patterns for primary care, which you can download.
    http://td.dartmouth.edu/centers/health-policy-research/pcsa/
    The instrument would be the percentage of CA ACEIs used in the PCSA of residence for each patient
    Problems
    Not condition specific. What if not very many people with the condition under consideration lived in a given beneficiary’s PCSA?

Creating instrument

- Approaches to defining the geographic area.

  - Create your own area out of concentric circles around ZIP codes
    Draw circles of various diameters around each ZIP code where a beneficiary lives.
    Problems
    Not condition specific. What if not very many people with the condition under consideration lived within X miles of a beneficiary’s ZIP code.
    Not appropriate for both rural and urban areas.
    Distances are arbitrary

Creating instrument

- Approaches to defining the geographic area.

  - Create your own using Driving Area of Clinical Care (DACC) (Fang, 2010)
    - Start with a given ZIP code.
    - Repeatedly add to it the next nearest ZIP code until you find X patients with the condition you are looking for.
    - X is somewhat arbitrary: Fang et al suggest starting at 10.
    - Once you get to X, this is the market area for the given ZIP code.
    - For that ZIP code, calculate the Area Treatment Ratio (ATR)
    - ATR’s greater than 1.0 represent “high intensity” areas.

Fang, L, Brooks, J, Chrischilles, E, "A New Method to Isolate Local Area Practice Styles as Prescribers Use as the Basis for Instrumental Variables in Comparative Effectiveness Research" Medical Care 48(8) (2010).
Creating instrument

• Geographic based instrument: DACC

  ▪ First step: calculate distances between ZIP codes
    GIS software works best because you can define the maximum driving time you will accept.
    Haversine formula (right) applied to latitude and longitude of ZIP code centroids.
    Latitudes and longitudes of ZIP codes can be downloaded from the US Census.

  ▪ Second Step: aggregate ZIP codes into DACCs (right)

  ▪ Third step: estimate logit and get predicted probability (see handout)

  ▪ Fourth step: Calculate ATR as shown on previous page (and handout)

---

Observed use and ATR

- Observed geographic distribution of high/low use of treatment ZIP codes (blue=high)
- High/low instrument ZIP codes after DACC adjustment

---

Overview

• Creating a geographic-based instrument
• Testing the correlation with treatment
• Test the correlation with observed confounders
• Getting the IV estimate
  - Two Stage Least Squares (2SLS)
  - Two-stage Residual Inclusion (2SRI)
  - Bivariate probit (BP)
• Repeat with provider preference instrument (on your own)
Correlation with treatment

- Looking for monotonicity and the 2-digit rule of thumb
- Divide the ATR into quartiles
- Look at the mean of the treatment across quartiles (see handout)
  It should increase more-or-less uniformly
- Run a first stage OLS of
  \[ T_i = \beta_i + \beta_2 X_i + \beta_3 Z_i + u_i, \]
  where \( Z \) is a vector of the three (of four) dummy variables for quartile of the instrument
- An F-test on the instruments should be at least 2 digits (i.e., \( \geq 10 \))

\[ \text{Correlation with treatment} \]

<table>
<thead>
<tr>
<th>Quartile of the DACC instrument</th>
<th>% of patients receiving treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-lowest</td>
<td>56.6%</td>
</tr>
<tr>
<td>1</td>
<td>70.9%</td>
</tr>
<tr>
<td>2</td>
<td>75.5%</td>
</tr>
<tr>
<td>3-highest</td>
<td>80.8%</td>
</tr>
</tbody>
</table>

Correlation with treatment

- Look for a continuous, monotonic increase in treatment with the value of the instrument.

- Perform an F-test on the instrument and hope for an F-statistic that is \( \geq 10 \).
Overview

- Creating a geographic-based instrument
- Testing the correlation with treatment
- Test the correlation with observed confounders
- Getting the IV estimate
  - Two Stage Least Squares (2SLS)
  - Two-stage Residual Inclusion (2SRI)
  - Bivariate probit (BP)

Correlation with observables

- Look for reduction in Mahalanobis distance

  Although this step is not technically necessary, it is politically necessary.
  - Your instrument can be correlated with observed variables, so long you include those variables in your model.
  - However, your story—that the instrument is uncorrelated with things you can’t observe—isn’t as believable if your instrument is correlated with things you can observe.
  - To do this compare the means of the covariates between patients with high and low values of the instrument, respectively.
  - Compare these differences in means with differences in the means of these covariates between patients who did and did not receive the treatment.

Correlation with observables

- Look for reduction in Mahalanobis distance

  Multiple testing and large sample sizes imply that t-tests and chi2 tests will probably be uninformative.
  - Instead, look for reductions in the distance between the vectors as a whole.
  - This can be done using Mahalanobis distance.
  - This is like Euclidean distance, i.e., $\sqrt{(a^2+b^2)}$ but it takes into account the variances and covariance(s) of the X vectors.

$$D = \sqrt{X' S X} \text{ where } S = X' X^{-1} X' \text{ and } X' \text{ indicates the transpose operator.}$$

Source: www.a1access.net
Correlation with observables

- Look for reduction in Mahalanobis distance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Count 1</th>
<th>Count 2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>10440</td>
<td>27646</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>75.8(7.33)</td>
<td>75.8(7.27)</td>
<td>0.471</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>68.1(7109)</td>
<td>66.4(18355)</td>
<td>0.002</td>
</tr>
<tr>
<td>Race (Black)</td>
<td>8.83(922)</td>
<td>7.07(1955)</td>
<td>0.000</td>
</tr>
<tr>
<td>Race (Asian/Pacific Island)</td>
<td>3.83(400)</td>
<td>2.58(712)</td>
<td>0.000</td>
</tr>
<tr>
<td>Race (Hispanic)</td>
<td>8.94(933)</td>
<td>5.87(1622)</td>
<td>0.000</td>
</tr>
<tr>
<td>Race (American Indian)</td>
<td>0.345(36)</td>
<td>0.770(213)</td>
<td>0.000</td>
</tr>
<tr>
<td>Race (Other Race)</td>
<td>0.929(97)</td>
<td>0.608(168)</td>
<td>0.001</td>
</tr>
<tr>
<td>Medicaid (Buy-in)</td>
<td>28.4(2969)</td>
<td>24.8(6847)</td>
<td>0.000</td>
</tr>
<tr>
<td>Log (Med HH Income)</td>
<td>10.6(0.353)</td>
<td>10.6(0.342)</td>
<td>0.072</td>
</tr>
<tr>
<td>Zip Code (Below Poverty)</td>
<td>0.131(0.086)</td>
<td>0.123(0.082)</td>
<td>0.000</td>
</tr>
<tr>
<td>Zip Code (&lt; HS Education)</td>
<td>0.302(0.090)</td>
<td>0.308(0.090)</td>
<td>0.000</td>
</tr>
<tr>
<td>Outpatient Visits Prior Year</td>
<td>9.5(8.21)</td>
<td>9.6(7.99)</td>
<td>0.330</td>
</tr>
<tr>
<td>Log (Med Cost Prior Year)</td>
<td>7.30(1.96)</td>
<td>7.58(1.96)</td>
<td>0.000</td>
</tr>
<tr>
<td>RXs Prior 90 Days</td>
<td>6.39(3.63)</td>
<td>6.74(3.80)</td>
<td>0.000</td>
</tr>
<tr>
<td>Hospitalized Prior Year</td>
<td>0.387(0.880)</td>
<td>0.535(1.04)</td>
<td>0.000</td>
</tr>
<tr>
<td>Flu Shot Prior Year</td>
<td>45.0(4698)</td>
<td>45.3(12530)</td>
<td>0.572</td>
</tr>
<tr>
<td>Elixhauser Score</td>
<td>4.67(7.00)</td>
<td>5.69(7.78)</td>
<td>0.000</td>
</tr>
<tr>
<td>Charlson Score</td>
<td>2.28(2.55)</td>
<td>2.54(2.73)</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Mahalanobis distance = 31.7

Compare this to 37.4 for unadjusted Mahalanobis distance: IV reduced distance by 45%

Overview

- Creating a geographic-based instrument
- Testing the correlation with treatment
- Test the correlation with observed confounders
- Getting the IV estimate
  - Two Stage Least Squares (2SLS)
  - Two-stage Residual Inclusion (2SRI)
  - Bivariate probit (BP)
Getting the IV estimate

**Two-stage least squares**
- Estimate OLS on treatment as a function of patient characteristics and instruments
- Predict treatment
- Use predicted treatment as regressor in OLS of outcome on predicted treatment and patient characteristics
- Better: Use ivreg2 in Stata. It will give you the correct standard errors

```plaintext
* 2SLS regression
xi: ivregress 2sls incidentalzat1yr $rhs / (treatment= i.iv_p_20_cat), first
2SRI
xi: reg treatment $rhs i.iv_p_20_cat
capture drop tx_hat
predict tx_hat, xb
xi: reg incidentalzat1yr tx_hat $rhs
```

**Two-stage residual inclusion**
- Estimate logit on treatment as a function of patient characteristics and instruments
- Predict treatment
- Calculate the residual
- Estimate a logit of the outcome as a function of observed treatment and the first stage residual.

```plaintext
* 2SRI
xi: logit treatment $rhs i.iv_p_20_cat
capture drop tx_hat
predict tx_hat, xb
xi: logit incidentalzat1yr treatment tx_hat $rhs
```

**Bivariate Probit**
- Estimate bivariate probit models.
- Simulations suggest bivariate probits are more efficient than 2SRI when you’ve got a dummy treatment and a dummy outcome.
- My simulations suggest bivariate probit is not as efficient as 2SRI, but these are preliminary results
- Use the biprobit command in Stata.
- Can take a while to converge. Keeping all X’s on a similar scale (e.g., transform income to log (income), divide age by a constant, etc.) can help.

```plaintext
* Bivariate probit
xi: biprobit (treatment= i.iv_p_20_cat $rhs) ///
    (incidentalzat1yr=treatment $rhs)
```
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Change in the probability of the outcome attributable to treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit unadjusted</td>
<td>2.6 (95% CI, 2.0 to 3.3)</td>
</tr>
<tr>
<td>Logit adjusted</td>
<td>1.8 (95% CI, 1.2 to 2.4)</td>
</tr>
<tr>
<td>2SLS</td>
<td>4.1 (95% CI, -6.6 to 14.9)</td>
</tr>
<tr>
<td>2SRI</td>
<td>2.2 (95% CI, -6.1 to 10.6)</td>
</tr>
<tr>
<td>Bivariate probit</td>
<td>0.6 (95% CI, -5.6 to 6.8)</td>
</tr>
</tbody>
</table>

Overview

- Creating a geographic-based instrument
- Testing the correlation with treatment
- Test the correlation with observed confounders
- Getting the IV estimate
  - Two Stage Least Squares (2SLS)
  - Two-stage Residual Inclusion (2SRI)
  - Bivariate probit (BP)

References & Resources