Canonical Correlation Analysis

TARDIS 2014

Celia M. Wilson, Ph.D.
Early Childhood Longitudinal Study-K

- 1998-99 kindergarten students, followed through 8th grade
  - Public and private schools
  - Full day and part day kinder programs
  - Diverse backgrounds
  - Children, parents, teachers, and schools
  - Children’s cognitive, social, emotional, and physical development
- Home, school, and classroom environment, class curriculum, teacher qualifications
Early Childhood Longitudinal Study-K

- Fall and spring of kinder (1998-99)
- Fall and spring of first grade (1999-2000)
  - Spring of third grade (2002)
  - Spring of fifth grade (2004)
  - Spring of eighth grade (2007)
Early Childhood Longitudinal Study-K

- Institute of Education Science’s National Center for Educational Statistics
- Nationally representative sample of kindergarteners, their parents, teachers, and schools across the United States
  - ECLS-B (born in 2001, tracked through 2007)
  - ECLS-K(2010-2011)
ECLS-K Variables (for this CCA)

- Gender (P1-fall of kinder)
- Age in months at start of kindergarten (P1-fall of kinder)
- Self Control Measure (P1-fall of kinder)
- Math T Score (C5-spring of 3rd grade)
- Reading T Score (C5-spring of 3rd grade)
# ECLS-K and CCA

<table>
<thead>
<tr>
<th>Independent Variables (Student Characteristics)</th>
<th>CCA</th>
<th>Dependent Variables (Student Achievement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Are these two variable sets related? If so, how?</td>
<td>Math score</td>
</tr>
<tr>
<td>Age in months</td>
<td></td>
<td>Reading score</td>
</tr>
<tr>
<td>Self control</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Are these two variable sets related? If so, how?
What is the goal of canonical correlation analysis (CCA)?

• To analyze the relationship between two sets of variables.

• Each subject is measured on two sets of variables.

• The researcher wants to know if and how the two sets are related to each other.
What happens during a CCA?

- Two sets of variables, each with at least two variables in each set.
- Variables in each set are linearly combined, maximizing shared variance, producing 2 canonical variates (two synthetic variables, one for each set).
- Canonical variates are then correlated, canonical correlation.
- The number of canonical correlations produced will equal the number of variables in the smaller variable set.
Canonical Correlation Analysis

First Canonical Variate $X$

$X_1 \rightarrow a_{x1}$
$X_2 \rightarrow a_{x2}$
$X_3 \rightarrow a_{x3}$

$R_{c1}$

First Canonical Variate $Y$

$Y_1 \rightarrow a_{y1}$
$Y_2 \rightarrow a_{y2}$

Synthetic variable $X$

Synthetic variable $Y$

$Age \rightarrow a_{x1}$
$Gend \rightarrow a_{x2}$
$SC \rightarrow a_{x3}$

$R_{c1}$

$Math \rightarrow a_{y1}$
$Read \rightarrow a_{y2}$
Data Considerations

- Ratio of cases to variables
- Normality, multivariate normality
  - Linearity
  - Homoscedasticity
  - Missing data
  - Absence of Outliers
- Absence of multicollinearity and singularity
ECLS-K and CCA

1. Descriptive statistics for each variable.
2. Evaluate data considerations and assumptions.
3. Run CCA.
4. Evaluate statistical significance.
5. Report effect sizes.
6. Provide a qualitative interpretation of findings/conclusion.
1. Descriptive statistics for each variable

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALE</td>
<td>3969</td>
<td>50.0</td>
</tr>
<tr>
<td>FEMALE</td>
<td>3973</td>
<td>50.0</td>
</tr>
<tr>
<td>Total</td>
<td>7942</td>
<td>100.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Std. Error</td>
</tr>
<tr>
<td>AGE(MNTHS)</td>
<td>7942</td>
<td>68.451</td>
<td>4.238</td>
<td>.180</td>
<td>-.503</td>
</tr>
<tr>
<td>READING T-SCORE</td>
<td>7942</td>
<td>51.818</td>
<td>9.733</td>
<td>-.061</td>
<td>-.293</td>
</tr>
<tr>
<td>MATH T-SCORE</td>
<td>7942</td>
<td>51.804</td>
<td>9.561</td>
<td>-.125</td>
<td>-.348</td>
</tr>
<tr>
<td>SELF-CONTROL</td>
<td>7942</td>
<td>2.8578</td>
<td>.489</td>
<td>-.613</td>
<td>1.02</td>
</tr>
<tr>
<td>Valid N</td>
<td>7942</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Evaluate data considerations and assumptions

- Ratio of cases to variables
- Normality, multivariate normality
  - Linearity
  - Homoscedasticity
- Missing data
- Absence of Outliers
- Absence of multicollinearity and singularity
3. Run CCA.

SPSS Syntax:
MANOVA
GENDER R1_KAGE SelfControl WITH ReadingT MathT
/PRINT=SIGNIF(MULTIV EIGEN DIMENR)
/DISCRIM=(STAN ESTIM COR ALPHA (.999)).
4./5. Evaluate statistical significance and report effect sizes

**Full canonical model**: evaluates the shared variance between the two sets of variables, across all the canonical functions.

Wilks lambda ($\lambda$) of .937, $F(6,15874)=87.515$, $p<.000$

Effect size: $1 - \lambda = .063$ (6.3% shared variance between the two variable sets, across all functions)
4./5. Evaluate statistical significance and report effect sizes

Each function separately:

Function 1: $R_{c1} = .209$, $R_{c1}^2 = .044$ (4.4%)

Function 2: $R_{c2} = .142$, $R_{c2}^2 = .020$ (2.0%)
4./5. Evaluate statistical significance and report effect sizes

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef</th>
<th>$r_s$</th>
<th>$r_s^2$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.166</td>
<td>.080</td>
<td>.006 (.6%)</td>
</tr>
<tr>
<td>Gender</td>
<td>1.000</td>
<td>.986</td>
<td>.972 (97.2%)</td>
</tr>
<tr>
<td>Self control</td>
<td>-.012</td>
<td>.022</td>
<td>.000 (0%)</td>
</tr>
<tr>
<td>Math</td>
<td>-1.26</td>
<td>-.290</td>
<td>.084 (8.4%)</td>
</tr>
<tr>
<td>Reading</td>
<td>1.36</td>
<td>.467</td>
<td>.218 (21.8%)</td>
</tr>
</tbody>
</table>
6. Interpretation of findings

For the 4.4% of shared variance between the two sets of variables for function 1, gender (97.2%), on the student characteristics side, seems to be contributing most. On the student achievement side, reading seems to be the most relevant variable at 21.8%.
References


