Correlation and Causation

Quantitative Methods for
Public Policy

Pete Ferderer
Associate Professor of Economics

February 24, 2005
Outline

• What is Causation?
• What is Correlation?
• Spurious Correlations
• Simpson’s Paradox
• How scientists determine causation
"Who was first?"
What is Causation?

When changes in one variable (X) affect changes in another variable (Y), we say that $X$ causes $Y$.

Examples:

Sun Rises $\rightarrow$ Rooster Crows (unidirectional)

Education $\rightarrow$ Higher wages (bidirectional?)
Important Questions

What causes poverty?
Did the Bush Tax cuts cause the economy to expand?
Does immigration cause lower wages?
Does Prozac cause suicide?
Does the burning of fossil fuel cause global warming?

Has the curse of the Bambino caused the Red Sox futility?
The Importance of Causality

*I would rather discover one causal law than be the King of Persia*

Democritus (460-370 B.C.)
How do we determine Causation?

• Correlation
• Controlled Experiments
• Theory
What is Correlation?

When two variables move together, we say they are correlated.
### Cross-Tabulation of 928 Adult Children Born of 205 Midparents, Sorted by Their Height and Their Midparent’s Height

<table>
<thead>
<tr>
<th>Height of Mid-parents (inches)</th>
<th>Height of the Adult Child</th>
<th>Total No. of Adult Children</th>
<th>Total No. of Mid-parents</th>
<th>Medians</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;61.7</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>66.3</td>
</tr>
<tr>
<td>62.2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>67.8</td>
</tr>
<tr>
<td>63.2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>67.9</td>
</tr>
<tr>
<td>64.2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>67.7</td>
</tr>
<tr>
<td>65.2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>67.9</td>
</tr>
<tr>
<td>66.2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>68.3</td>
</tr>
<tr>
<td>67.2</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>68.5</td>
</tr>
<tr>
<td>68.2</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>69.0</td>
</tr>
<tr>
<td>69.2</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>70.0</td>
</tr>
<tr>
<td>70.2</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>71.2</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>72.2</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>73.2</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>&gt;75.7</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

How do we measure Correlation?

Covariance = \( \frac{1}{3} \left[ \sum_{i=1}^{3} (X_i - \bar{X})(Y_i - \bar{Y}) \right] \)

= \( \frac{1}{3} \left[ (-1)(-1) + (0)(0) + (1)(1) \right] \)

= \( \frac{2}{3} \)
How do we measure Correlation?

\[
\text{Covariance} = \frac{1}{3} \left[ \sum_{i=1}^{3} (X_i - \bar{X})(Y_i - \bar{Y}) \right]
\]

\[
= \frac{1}{3} \left[ (-1)(1) + (0)(0) + (1)(-1) \right] = -\frac{2}{3}
\]
How do we measure Correlation?

Covariance = \( \frac{1}{5} \left[ \sum_{i=1}^{5} (X_i - \bar{X})(Y_i - \bar{Y}) \right] \)

= \( \frac{1}{5} \left[ (-1)(-1) + (0)(0) + (1)(1) + (-1)(1) + (1)(-1) \right] = 0 \)
How do we measure Correlation?

The correlation coefficient is a measure of association, which is normalized

\[ r = \text{correlation Coefficient} \]

\[ r = \frac{\text{Covariance}(X, Y)}{\sqrt{\text{Variance}(X)} \sqrt{\text{Variance}(X)}} \]

First Example

\[ r = \frac{2/3}{\sqrt{2/3} \sqrt{2/3}} = 1 \]

\[-1 \leq r \leq +1 \quad \text{No correlation: } r = 0\]
Which Way does Causality Run?

Source: Macroeconomics, by Greg Mankiw

$r = 0.81$
Which Way does Causality Run?

Source: *Macroeconomics*, by Greg Mankiw

\[ r = -0.68 \]
Crime and Punishment

Crime versus In Prison:  \( r = 0.54905 \)
Crime versus poverty:    \( r = 0.25187 \)
Crime versus HS dropout: \( r = 0.51496 \)

What causes what?
Are correlations “statistically significant?”
(Professor Miao’s talk)

http://www.huppi.com/kangaroo/L-toughcrime.htm
Crucial Point:

CORRELATION DOES NOT NECESSARILY IMPLY CAUSALITY
Example 1:
Teachers’ Salary and SAT Scores
Example 1:
SAT score versus Fraction Taking Test
Example 1:
Salary Versus Fraction Taking Test

% Taking

Salary
Example 1:
Possible Causal Relationships

↑ Salary
↓ SAT Scores
↑ % Taking Exam

Spurious!

How do we know?
Important Point

The correlation may be merely a coincidence.
The correlated effects may have a common underlying cause.
One of the correlated effects may be the cause of the other.
Example 2

Researchers found a correlation of 0.86 between the number of churchgoers and the number of burglaries committed in different towns.

Explanation?

Attending church makes people want to rob.

More churchgoers means more empty houses.

Common Third Cause: Population
Sir Francis Galton: A Cautionary tail

Founder of “eugenics”

Belief: talent was based on heredity alone

Evidence: strong positive correlation between talent of parents and offspring (e.g., judges had children that were judges)

Policy Goal: Limit reproduction of less talented or ill
Galton Society Established in the U.S. in 1918

Goal: Undo U.S Immigration Policy

Evidence: Immigrants who had been in the U.S. longer scored higher on I.Q. tests than recent immigrants

Interpretation: Flow of recent immigrants is increasingly stupid.
Simpson’s Paradox

A Statistical relationship between two variables can be reversed by including additional factors in the analysis.

Occurs when we combine smaller data sets into larger ones.

Example of Simpson’s Paradox

Average College Physics Grade for students in Engineering Program

<table>
<thead>
<tr>
<th></th>
<th>HS Physics</th>
<th>None</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>50</td>
<td>5</td>
<td>------</td>
</tr>
<tr>
<td>Ave Grade</td>
<td>80</td>
<td>70</td>
<td>10</td>
</tr>
</tbody>
</table>
Example of Simpson’s Paradox

Average College Physics Grade for students in Liberal Arts Program

<table>
<thead>
<tr>
<th></th>
<th>HS Physics</th>
<th>None</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>5</td>
<td>50</td>
<td>----</td>
</tr>
<tr>
<td>Ave Grade</td>
<td>95</td>
<td>85</td>
<td>10</td>
</tr>
</tbody>
</table>

Conclusion: In both classes taking high school physics improved the grade
## Example of Simpson’s Paradox

Average College Physics Grade for students who took high school physics

<table>
<thead>
<tr>
<th></th>
<th># Students</th>
<th>Grades</th>
<th>Grade Pts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td>50</td>
<td>80</td>
<td>4000</td>
</tr>
<tr>
<td>Liberal Arts</td>
<td>5</td>
<td>95</td>
<td>475</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>55</strong></td>
<td></td>
<td><strong>4475</strong></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>----</td>
<td><strong>81.4</strong></td>
<td>----</td>
</tr>
</tbody>
</table>
Example of Simpson’s Paradox

Average College Physics Grade for students who did not take high school physics

<table>
<thead>
<tr>
<th># Students</th>
<th>Grades</th>
<th>Grade Pts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>Liberal Arts</td>
<td>50</td>
<td>85</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

Average 83.6
Example of Simpson’s Paradox

In both college physics classes taking high school physics improved grade

Yet those who had taken high school physics had lower grades in college physics courses

Why?

Lurking Variable: Course difficulty
How do we determine causality?

Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).

Albert Einstein (1953)
Uncontrolled Conditions

Socio-economic status
(e.g., diet, life style)

Treatment

Recovery

From Pearl *Causality* (2000)
Uncontrolled Conditions

Taking treatment is correlated with recovery, not because treatment causes recovery, but because the people who take the treatment are also those who live healthier lifestyle—the true cause of recovery.

Thus the correlation between treatment and recovery is spurious.
Experimental Conditions

- Coin
- Socio-economic status (e.g., diet, lifestyle, age)
- Treatment
- Recovery
How do Social Scientists Determine Causality?

• Theory

• Natural Experiments
Why are we so sure that....

Sun Rises $\rightarrow$ Rooster crows (unidirectional)

We understand the forces which make the sun rise and the Rooster react to light
Natural Experiment

Donohue and Levitt (2001) QJE

Claim: Large and unexpected drop in crime rates in 1990s caused by abortion legalization in 1970s.

unwanted children more likely to grow up as criminals

unwanted children more likely to be aborted

Natural Experiment: Crime rates fell first in states that loosened laws first.
Example 4: Immigration & Wages

Pennies per Hour

% Foreign Born

0 .05 .1 .15 .2 .25 .3 .35

1000 1100 1200 1300 1400 1500

Maine New Hampshire Vermont Massachusetts
Virginia Washington Maryland
Connecticut New Jersey Delaware
Pennsylvania New York Maryland
Colorado Idaho Wyoming North Dakota
Montana Nevada Utah Idaho
Arizona Nevada Hawaii Washington
Oregon California Alaska Hawaii