Correlation Coefficients: Stepping Stones to Impactful Insights

Kuba Glazek, Ph.D.
Methodology Expert
National Center for Academic and Dissertation Excellence Excellence
Los Angeles Campus
Letters in Winning Word of Scripps National Spelling Bee correlates with Number of people killed by venomous spiders

Correlation: 80.57% (r=0.8057)

Data sources: National Spelling Bee and Centers for Disease Control & Prevention (tylervigen.com)
Outline

• Overview of the correlation coefficient
• The trouble with correlations
  – Third variables
  – Coincidence
  – Reverse-causation
• Solutions
• Examples
Correlation Overview

• Pearson’s $r$: Statistically speaking, does a linear relationship exist between two continuous variables?

• Three types of linear relationships among variables are possible
  – As one increases, the other increases (positive correlation)
  – As one increases, the other decreases (negative correlation)
  – As one increases, the other does nothing (no correlation)

• Strength of correlation
  – 0 to 1 (positive correlations)
  – 0 to -1 (negative correlations)
Correlations are Useful

• Preliminary analyses
  – Establish that there is no lack of relationship
  – Describe strength of relationship

• Establish test-retest reliability
  – E.g., in scale development, you want to know that performance is consistent over time

• Underlying technique of lots of other analyses
  – T-tests
  – Regression
So What’s the Problem?

- Spurious correlation
  - Third variable
  - Coincidence
- Reverse causation
“Now go to sleep, Kevin—or once again I’ll have to knock three times and summon the Floating Head of Death.”
Third Variable Problem

- Sleeping with one’s shoes on is positively correlated with waking up with a headache. Therefore, sleeping with one’s shoes on causes headaches.
- As ice cream sales increase, the rate of drowning deaths increases. Therefore, eating ice cream causes drowning.
- Amount of art training is significantly correlated with math SAT scores (Hetland & Winner, 2000)
Coincidence

Letters in Winning Word of Scripps National Spelling Bee correlates with Number of people killed by venomous spiders

Correlation: 80.57% (r=0.8057)

Data sources: National Spelling Bee and Centers for Disease Control & Prevention (tylervigen.com)
Reverse Causation

• Direction of causality other than intuited
  – The more firefighters fighting a fire, the bigger the fire. Therefore, firemen cause fires.
  – The more sign language, the faster verbal vocabulary develops (Acredolo & Goodwyn, 1988). Therefore, sign language causes verbal language.
Logic for Extending Correlational Findings

• Causal claims based on correlational models rest on the following assumptions (e.g., Preacher, Rucker, & Hayes, 2007, p. 217)
  – Predictor variable precedes outcome variable in time (logic, knowledge base, etc.)
  – Concomitant variation of predictor and outcome variable (knowledge base, data)
  – *Sources of spurious correlation have been eliminated* (statistical analysis)
Controlling Spurious Correlations

• Those who become fathers at younger ages die younger (i.e., age at fatherhood positively correlated with age at death) (Einiö, Nisén, & Martikainen, 2015)

  ✓ Predictor variable precedes outcome variable in time (becoming a father precedes death)

  ✓ Concomitant variation of predictor and outcome variable (data confirm correlation)

  – Sources of spurious correlation have been eliminated?

Review the Literature

• Any potential third factors?
• Educational level was controlled, because of its close links with both fertility timing and mortality (16, 23) …
• We further controlled for marital status, because young fatherhood was associated with unstable marriages in our sample, and marriage has been shown to be associated with decreased mortality (25).
• The number of children was controlled, because those who have a child early in life tend to have more children, and it has been suggested, it increases all-cause mortality—at least among women (26–28).
• Region of residence was taken into account, because of its close links with both fertility timing and mortality in Finland (29, 30). (Einiö et al., 2015, p. 2)
Review the literature

• Any potential third factors?
• Educational level was controlled, because of its close links with both fertility timing and mortality (16, 23) ...
• We further controlled for marital status, because young fatherhood was associated with unstable marriages in our sample, and marriage has been shown to be associated with decreased mortality (25).
• The number of children was controlled, because those who have a child early in life tend to have more children, and it has been suggested, it increases all-cause mortality—at least among women (26–28).
• Region of residence was taken into account, because of its close links with both fertility timing and mortality in Finland (29, 30). (Einiö et al., 2015, p. 2)
Two Variables

IV

DV
Correlation Between Two Variables
Third Variable (Covariate)
Controlling for Nuisance Variable
• Purple shaded area represents variance in DV accounted for by covariate

• Note decrease in extent of green area (correlation between IV and DV) when covariate is included

• If covariate “overshadows” green area, there is no reason to believe a special/preferential relationship exists between IV and DV (unless theory dictates that covariate mediates relationship between IV and DV)
A Different Example
Controlling Covariate Leaves Relatively Large Correlation Intact

$r$
• Purple shaded area represents variance in IV-DV relationship accounted for by covariate
  – Does not “overshadow” green area
  – There is a reason to believe a special/preferential relationship exists between IV and DV
• Included covariate was well-motivated (i.e., based on literature, logic, etc.) but predicted relationship exists even after controlling for it!
• Of course, what other covariates are out there???
Choose Control Variables Wisely

• Limit number of control variables
  – In multiple regression, need ratio of at least 15 participants per variable (e.g., if analyzing 3 variables, need at least N = 45)
  – In ANCOVA, should have no more than 2 or 3 control variables (a.k.a. covariates)

• Types of control variables
  – In multiple regression, can be categorical or continuous
  – In ANCOVA, need to be continuous (if categorical, covariate becomes a second IV)
Conducting the Analysis in SPSS
Multiple Regression

• Can be linear regression (i.e., continuous DV)
• Can be logistic regression (i.e., categorical DV)

• Can use continuous and/or categorical IVs and covariates
A Bare-Bones Example Using SPSS

• Do stressful life events predict number of visits to a health professional?
• Literature review and professional experience suggest that physical and mental health symptoms may be related to number of visits
• All variables in example are continuous
• Dataset available at
  http://wps.ablongman.com/ab_tabachnick_multistats_6/

a: See NCADE YouTube channel videos: “Making sense of quantitative data” (parts 1 and 2).
## Correlations in Data

<table>
<thead>
<tr>
<th></th>
<th>Correlations</th>
<th>Stressful life events</th>
<th>Mental health symptoms</th>
<th>Physical health symptoms</th>
<th>Visits to health professionals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stressful life events</td>
<td>Pearson Correlation</td>
<td>1</td>
<td>.306**</td>
<td>.370**</td>
<td>.287**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>.000</td>
<td>1</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>465</td>
<td>465</td>
<td>465</td>
<td>465</td>
</tr>
<tr>
<td>Mental health symptoms</td>
<td>Pearson Correlation</td>
<td>1</td>
<td>1</td>
<td>.505**</td>
<td>.440**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>.000</td>
<td>.000</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>465</td>
<td>465</td>
<td>465</td>
<td>465</td>
</tr>
<tr>
<td>Physical health symptoms</td>
<td>Pearson Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>465</td>
</tr>
<tr>
<td>Visits to health professionals</td>
<td>Pearson Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>465</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
**Conclusion:** Stress accounts for a significant amount of variance in number of doctors visits (2.4%) after controlling for physical and mental health.
Conclusion: Mental health symptoms are the strongest predictor of doctor visits, followed by stressful life events.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>95.0% Confidence Interval for B</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
<td>Sig.</td>
<td>Lower Bound</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>-2.319</td>
<td>1.077</td>
<td>-2.152</td>
<td>.032</td>
<td>-4.436</td>
<td>-.202</td>
</tr>
<tr>
<td>Physical health symptoms</td>
<td>.118</td>
<td>.126</td>
<td>.045</td>
<td>.933</td>
<td>.351</td>
<td>-.130</td>
</tr>
<tr>
<td>Mental health symptoms</td>
<td>1.910</td>
<td>.222</td>
<td>.417</td>
<td>8.616</td>
<td>.000</td>
<td>1.475</td>
</tr>
<tr>
<td>2 (Constant)</td>
<td>-3.705</td>
<td>1.124</td>
<td>-3.296</td>
<td>.001</td>
<td>-5.914</td>
<td>-1.496</td>
</tr>
<tr>
<td>Physical health symptoms</td>
<td>-.010</td>
<td>.129</td>
<td>-.004</td>
<td>-.075</td>
<td>.940</td>
<td>-.263</td>
</tr>
<tr>
<td>Mental health symptoms</td>
<td>1.787</td>
<td>.221</td>
<td>.390</td>
<td>8.083</td>
<td>.000</td>
<td>1.353</td>
</tr>
<tr>
<td>Stressful life events</td>
<td>.014</td>
<td>.004</td>
<td>.169</td>
<td>3.769</td>
<td>.000</td>
<td>.007</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Visits to health professionals
ANCOVA in a Nutshell

- IV (or IVs) are categorical
  - E.g., age group
  - E.g., gender
  - If a continuous IV violates the assumption of being normally distributed (bell curve), can be dichotomized into high/low groups (e.g., good physical health and poor physical health)

- Covariates are continuous
- DVs are continuous
- Same as analysis of variance (ANOVA), except with extra C (covariance)
ANCOVA in a Nutshell

ANCOVA in SPSS

Analyze → General Linear Model → Univariate →
Thank You!
Questions?

Presentation available on YouTube
• Search for “NCADE”

Email: ncade.me@thechicagoschool.edu
Office: Los Angeles campus, room 748
Phone: (213) 615-7290