Correlation & Regression

• Over the next two weeks, we discuss two concepts that focus on the relationship between variables: correlation and regression

• Correlation: a measure of the linear association between 2 variables

• Regression: estimates the “best-fitting” line that relates a predictor variable, X, to a criterion variable, Y.
  - used to understand how Y changes when X varies

Correlation of Variables

• Correlation can be calculated for:
  - dependent variable ($Y_i$) and independent variable ($X_i$)
  - 2 dependent variables ($Y_{1i}$ and $Y_{2i}$)
  - 2 independent variables ($X_{1i}$ and $X_{2i}$)

• Examples from hypothetical study examining antidepressants & sleep:
  - $Y_i$ = mood state, $X_i$ = antidepressant dose
  - $Y_{1i}$ = sleep quality, $Y_{2i}$ = time slept
  - $X_{1i}$ = antidepressant dose, $X_{2i}$ = type of depression

Correlations in Research

• effectiveness of fluoridated water in reducing cavities
• association between smoking and lung cancer
• apparently false claim of link between autism and measles, mumps, and rubella (MMR) inoculations – proposed that MMR causes autism
• drinking coffee associated with longer life
• claim that orchestra conductors live longer
Exploring data so far…

Correlation: Is there a relationship between these 2 variables?

Scatterplot

Correlation strength & direction

Correlation depends on fit, not slope
Perfect Linear Relation (r = 1 or r = -1)

- Perfect positive correlation
- Perfect negative correlation

No linear association (r = 0)

- Dotted lines indicate means of x and y.
- Note that points do not fall within each quadrant equally.

Correlations & deviations around the means

- r = 0.7

Linear vs. curvilinear relations

- **Monotonic**
  - Linear
  - Non-linear
- **Non-monotonic**
  - Non-linear

- **Monotonic**: as X increases, Y increases or decreases without reversal
  - might not be a straight line!
- **Non-monotonic**: as X increases, Y changes direction at least once
- **Linear** relationship: best fit line is straight
- **Curvilinear**: best fit “line” is not straight
- **Correlation** is only sensitive to linear relations!
4 \((x,y)\) data sets plotted with a reference line

Question: which set has the highest correlation?

The correlation is the same in all 4 sets! \(r = 0.82\)

Correlation is a measure of linear association and is insensitive to non-linear association.

Covariance

- Statistic representing the degree to which 2 variables vary together

\[
\text{cov}_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}
\]

- where N is the number of observations.

- Note that covariance is the sum of products of deviation scores

Green points drive \(\text{COV}_{XY}\) in positive direction.
Blue points drive \(\text{COV}_{XY}\) in negative direction.
Net \(\text{COV}_{XY}\) differs from zero.
Green points drive COV\(_{XY}\) in positive direction. Blue points drive COV\(_{XY}\) in negative direction. Note that net COV\(_{XY}\) is zero.

\[
\text{COV}_{XY} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}
\]

**Correlation Coefficient, r**

\[
r = \frac{\text{COV}_{XY}}{s_X s_Y}
\]

- Standardizes the covariance so that we can compare correlations even between groups with very different dispersions
- \(s_X\) and \(s_Y\) are the standard deviations of \(X\) and \(Y\) respectively.
- \(r\) is the Pearson Product-Moment Correlation Coefficient
- Note that \(r\) has no “units”

**Pearson r**

- \(r\) ranges from -1 to 1
  - \(r = +1\) when \(X\) and \(Y\) are perfectly positively correlated
  - \(r = -1\) when \(X\) and \(Y\) are perfectly negatively correlated
  - \(r = 0\) when there is no linear relationship between \(X\) and \(Y\)
- in Psychology,
  - Weak \(r\): 0.1-0.3
  - Moderate \(r\): 0.3-0.5
  - Strong \(r\): >0.5

**What does \(r\) mean?**

Measures magnitude (strength) & direction of linear relationship between 2 variables

- \(r = +1\)
- \(r = -1\)
- \(r = 0\)

Q2. But, this looks like a perfect correlation…?

\(r\) is an index of how much uncertainty about \(Y\) is reduced by knowing \(X\).
r varies across samples

Confidence Interval

- A confidence interval is a range of values, calculated from the sample observations, that are believed, with a particular probability, to contain the true population parameter. A 95% confidence interval, for example, implies that were the estimation process repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value. Note that the stated probability level refers to the properties of the interval and not to the parameter itself which is not considered a random variable. – B.S. Everitt, Dictionary of Statistics

- 95% Confidence Interval
  - a range (interval) for a statistic (e.g., r)
  - calculated from your sample
  - the interval varies across samples
  - in the long run, the interval contains the true population value 95% of the time
  - example: the sample correlation r was 0.37 (CI_{95%} = [-0.08, 0.70])

Other Types of Correlations

- Two continuous variables
  - Pearson Product-Moment Correlation Coefficient (r)
- Two ranked/ordinal variables
  - Spearman’s correlation coefficient for ranked data (rho (ρ) or \( r_s \))
- One dichotomous & one continuous variable
  - e.g., correct/incorrect answer on mc question and exam total score
  - Point-biserial Correlation (\( r_{pb} \)): Calculate Pearson’s r but call it \( r_{pb} \)
- Two dichotomous variables
  - e.g. age (teens/seniors) and coin purse ownership (yes/no)
  - Calculate Pearson’s r but called in \( r_{p} \)

Correlations with ranked data

- Spearman's correlation coefficient, \( r_s \)
  - useful when observations have been replaced by their numerical ranks
    - e.g., changing applicants’ exam scores to ranks

<table>
<thead>
<tr>
<th>english</th>
<th>math</th>
<th>english.rank</th>
<th>math.rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>66</td>
<td>2.0</td>
<td>7</td>
</tr>
<tr>
<td>75</td>
<td>70</td>
<td>8.0</td>
<td>9</td>
</tr>
<tr>
<td>45</td>
<td>40</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>71</td>
<td>60</td>
<td>7.0</td>
<td>4</td>
</tr>
<tr>
<td>61</td>
<td>65</td>
<td>4.5</td>
<td>6</td>
</tr>
<tr>
<td>64</td>
<td>56</td>
<td>6.0</td>
<td>2</td>
</tr>
<tr>
<td>58</td>
<td>59</td>
<td>3.0</td>
<td>3</td>
</tr>
<tr>
<td>80</td>
<td>77</td>
<td>10.0</td>
<td>10</td>
</tr>
<tr>
<td>76</td>
<td>67</td>
<td>9.0</td>
<td>8</td>
</tr>
<tr>
<td>61</td>
<td>63</td>
<td>4.5</td>
<td>5</td>
</tr>
</tbody>
</table>
Spearman’s rank order correlation

Same as Pearson r correlation for ranks, not actual (x,y) values. Why use $r_s$ if you have the original values?

$r_s$ is more robust to extreme “high-leverage” points

Changing only 2 out of 100 points markedly affects $r$ but not $r_s$

...but not to all types of extreme scores

Correlations with ranked data

- Spearman’s correlation coefficient, $r_s$, calculated with the same formula that is used to calculate Pearson’s $r$
  - formula applied to ranks, not actual (X,Y) values
- $r_s$ is much more resistant/robust than $r$ to high leverage data points
- However, interpretations of $r$ and $r_s$ differ:
  - $r$ is an index of the strength/direction of linear relation
  - $r_s$ is an index of the strength/direction of monotonic relation
$r_s$ measures monotonicity of $X,Y$ association

$r_s$ measures linear association between $X,Y$ ranks

$r_s$ measures monotonic association between $X,Y$ scores

$r_s$ is sensitive to extreme scores affecting monotonicity

Inclusion of Northern Ireland data point significantly alters monotonicity of the $X,Y$ association

**Other Types of Correlations**

- Two continuous variables
  - Pearson Product-Moment Correlation Coefficient ($r$)
- Two ranked/ordinal variables
  - Spearman's correlation coefficient for ranked data (rho ($\rho$) or $r_s$)
- One dichotomous & one continuous variable
  - e.g., correct/incorrect answer on mc question and exam total score
  - **Point-biserial Correlation ($r_{pb}$)**: Calculate Pearson's $r$ but call it $r_{pb}$
- Two dichotomous variables
  - e.g. age (teens/seniors) and coin purse ownership (yes/no)
  - Calculate Pearson's $r$ but called in $r_s$ (phi)

**Point-biserial Correlation $r_{pb}$ example**

- **X variable**: answers for a single multiple-choice question were coded as “incorrect” (0) or “correct” (1)
- **Y variable**: total score on remaining exam questions
- Correlate $X$ & $Y$: $r_{pb} = 0.46$
- What happens if we code “incorrect” and “correct” responses with other numbers (e.g., -10 & 10)?
  - magnitude $r_{pb}$ is not changed by coding scheme for $X$ variable (sign of $r_{pb}$ can change)
Other Types of Correlations

• Two continuous variables
  - Pearson Product-Moment Correlation Coefficient (r)

• Two ranked/ordinal variables
  - Spearman’s correlation coefficient for ranked data (rho (\( \rho \)) or \( r_s \))

• One dichotomous & one continuous variable
  - e.g., correct/incorrect answer on mc question and exam total score
  - Point-biserial Correlation (\( r_{pb} \)): Calculate Pearson’s r but call it \( r_{pb} \)

• Two dichotomous variables
  - e.g., gender (male/female) and correct/incorrect answer on mc question
  - Calculate Pearson’s r but called in \( r_\phi \) (phi)

Correlations with ranked data

- It makes sense to investigate the correlation between English (mark) and Maths (mark) using r, because the data is on a ratio scale and the intervals are meaningful.
- It makes sense to use rs for investigating the relationship between Rank (English) and Rank (Maths) because the data are ranked on an ordinal scale and the intervals are meaningless.

Factors that affect correlation

- Nonlinearity
- Extreme observations
- Range restrictions
- Heterogeneous subsamples
**Restricted Range**

- Restricted range can obscure a linear association.
- Restricted range can obscure a curvilinear association.

**Heterogeneous Subsamples**

- Simpson (1951):
  - A statistical association observed in a population could be attenuated and even reversed within subgroups that make up that population.
- Correlations calculated on populations consisting of heterogeneous subgroups may be misleading.

**Simpson’s Paradox**

- Correlation for population is positive.
- Correlations within sub-groups are negative.
Simpson’s Paradox

Overall correlation is approximately zero
Strong, opposite correlations in 2 sub-groups

Correlation ≠ Causation

- reading skill correlated with shoe size
- over last 2 centuries, price of bread in Great Britain is correlated with sea level in Venice
- \( r_{pb} \): BMW owners have higher incomes than Ford owners
- number of pirates is correlated with global temperature

Pirates Cause Global Warming

Correlation ≠ Causation cont’d

1. The relationship could be causal.
   - Sunlight could protect against breast cancer (because it helps produce vitamin D)
2. The relationship may be reversed.
   - Happiness could lead to more social relationships, or the other way around.
3. The relationship could be partly causal.
   - Increased wealth could lead to increased happiness (if a supportive family is also present)
4. Variables may be changing over time.
   - Pirates and global warming are associated because both change over time
5. There may be other confounding factors.
   - Reading skill correlated with shoe size because both are correlated with age
   - Wine consumption is associated with decrease in heart disease
   - Wine consumption is greater in sunny climates, and heart disease is lower in sunny regions
   - Perhaps wine/heart disease association really is sun-exposure/heart disease association?
6. Both may be other causal factors.
   - Reading skill correlated with shoe size because both correlated with years of education
   - Family stability and physical illness may be related because stress is related to both.
7. Correlation may be a coincidence (due to chance).

Correlations (summary)

- Correlations: index of the association between 2 variables
- Pearson r: index of strength & direction of linear association
  - \( r = \frac{COV_{XY}}{s_X s_Y} \); varies between -1 & +1
  - A measure of how much our uncertainty about the value of Y is reduced by knowing the value of X
  - \( r \) is an estimate of the population correlation that varies across samples
  - A 95% Confidence Interval of \( r \), which varies from sample to sample, contains the true population parameter 95% of the time
- Spearman’s \( r_s \): index of monotonicity of association
  - Equivalent to Pearson’s \( r \) for ranked data
  - Generally more robust than \( r \) to extreme (i.e., “high-leverage”) data points
- Correlations affected by: nonlinearity, extreme scores, range restrictions, heterogeneous samples
- Correlation ≠ Causation