Correlation

Jonathan A. Pedroza, PhD

Modeling Relationships

$$outcome_i = (model) + error_i$$

$$outcome_i = (b_1 X_i) + error_i$$

- if we work with standardized scores (what are those called?) then the equation changes because the predictor and outcome have a mean of 0

Modeling Relationships

- we would lose the intercept so what is left over is the following equation

$$z(outcome)_i = b_1 z(X_i) + error_i$$

- wit this equation the outcome can be presented as a standardized score predicted by a standardized variable multiplied by b_1

Modeling Relationships

- when using standardized scores, the value becomes a pearson product-moment correlation coefficient
  - pearson product-moment correlation coefficient = correlation coefficient
    - denoted with a r
    - which means the correlation coefficient or r is standardized

Covariance

- simply put, **covariance** is an un-standardized correlation
- to understand covariance, we must look at variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - X)^2}{N-1} = \frac{\sum_{i=1}^{n}(X_i - X)(X_i - X)}{N-1}$$

- covariance is the examination of the relationship between two variables
  - if one variable deviates from its mean, the other variable should either deviate from its mean in the same direction or in the opposite direction

Covariance

- for variance, we ________ our deviations
  - for covariance, we multiply the deviation of one variable by the deviation for the second variable
    - positive values indicate a relationship in the same direction
    - negative values indicate a relationship where the deviations are in opposite directions
- multiplying deviations of one variable by the deviations of a second variable provides **cross-product deviations**
- we then get the average by dividing by the degrees of freedom $(N-1)$

$$covariance(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{N-1}$$

Covariance Example

```
  x  y
1 1 10
2 4  9
3 5  8
4 6  6
5 7  8
```

Covariance Example

```r
mean(example$x)
```

```
[1] 4.6
```

```r
mean(example$y)
```

```
[1] 8.2
```

Covariance Example

```r
example$x_deviations <- example$x - 4.6
example$y_deviations <- example$y - 8.2

example
```

```
  x  y x_deviations y_deviations
1 1 10         -3.6          1.8
2 4  9         -0.6          0.8
3 5  8          0.4         -0.2
4 6  6          1.4         -2.2
5 7  8          2.4         -0.2
```

Covariance Example

$$covariance(x,y) = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{N-1}$$

Covariance Example

$$\frac{(1-4.6)(10-8.2)+(4-4.6)(9-8.2)+(5-4.6)(8-8.2)+(6-4.}{5-1}$$

Covariance Example

```r
(1 - 4.6)
```

```
[1] -3.6
```

```r
(10 - 8.2)
```

```
[1] 1.8
```

```r
(4 - 4.6)
```

```
[1] -0.6
```

```r
(9 - 8.2)
```

```
[1] 0.8
```

```r
(5 - 4.6)
```

```
[1] 0.4
```

Covariance Example

```r
(8 - 8.2)
```

```
[1] -0.2
```

```r
(6 - 4.6)
```

```
[1] 1.4
```

```r
(6 - 8.2)
```

```
[1] -2.2
```

Covariance Example

```r
(7 - 4.6)
```

```
[1] 2.4
```

```r
(8 - 8.2)
```

```
[1] -0.2
```

```r
5 - 1
```

```
[1] 4
```

Covariance Example

$$\frac{(-3.6)(1.8) + (-.6)(.8) + (.4)(-.2) + (1.4)(-2.2) + (2.4)(-.2)}{4}$$

Covariance Example

```
(-3.6)*(1.8)
```

```
[1] -6.48
```

```
(-.6)*(.8)
```

```
[1] -0.48
```

```
(.4)*(-.2)
```

```
[1] -0.08
```

```
(1.4)*(-2.2)
```

```
[1] -3.08
```

```
(2.4)*(-.2)
```

```
[1] -0.48
```

Covariance Example

$$\frac{(-6.48) + (-.48) + (-.08) + (-3.08) + (-.48)}{4}$$

Covariance Example

```r
(-6.48) + (-.48) + (-.08) + (-3.08) + (-.48)
```

```
[1] -10.6
```

$$covariance(x, y) = \frac{-10.6}{4}$$

Covariance Example

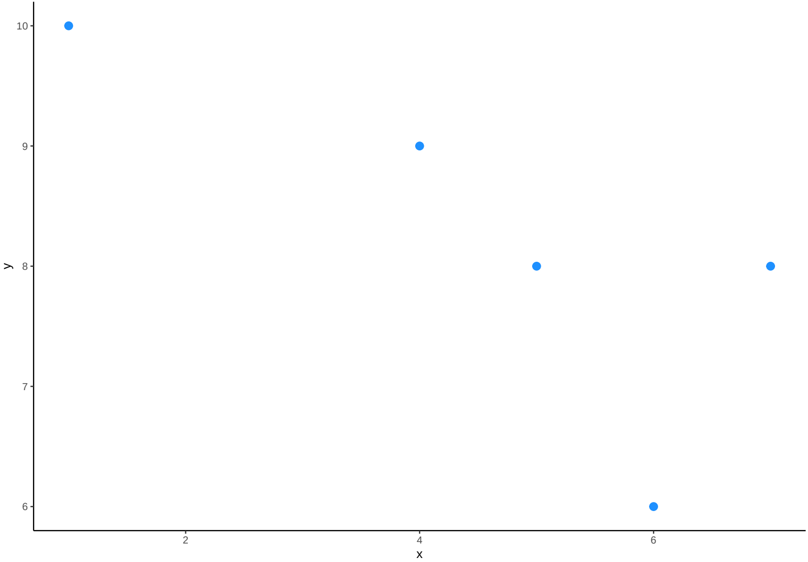
```
-10.6/4
```

```
[1] -2.65
```

Covariance Example

$$covariance(x, y) = -2.65$$

Covariance Example



Covariance Example

- since this is a negative covariance, which of the following would be true
  - both variables deviated from the mean in the same direction
  - one variable deviated away from the mean (increased) while one variable deviated from the mean in the opposite direction (decreased)

Standardization & Correlation Coefficients

- **standardization** is the process of converting the covariance into standardized units
  - the unit of measurement we are looking for are standard deviation units
    - **standard deviation** is the average deviation from the mean
- to standardize our covariance, we would divide by the standard deviation
  - we have 2 standard deviations though
  - just like with our deviations, we are going to multiply our standard deviations

Standardization & Correlation Coefficients

- so our covariance value is divided by the product of our multiplied standard deviations
  - this is known as a **correlation coefficient**

$$r = \frac{cov_{xy}}{S_x S_y} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{(N-1)S_x S_y}$$

- this correlation coefficient is the **Pearson product-moment correlation coefficient** or simply **Pearson's correlation coefficient** or r

Standardization & Correlation Coefficients

- by standardizing our covariance, we now can only have values that go from -1 to 1
  - these are seen by the strength of the relationship
    - r =0 −> no correlation
    - r = .1 is a small/weak correlation/effect size
    - r = .3 is a moderate/medium correlation/effect size
    - r = .5 is a large correlation/effect size

Standardization & Correlation Coefficients

- JP: while these values can go from -1 to 1, negative and positive values don't matter in regard to strength
  - r = -.8 is a larger correlation than r = .6
  - the larger number will always be the larger correlation
  - negative values only indicate direction
    - r = -.8 is a large negative/inverse correlation
- when we examine a correlation between two variables, we are using a **bivariate correlation**

Significance of Correlation Coefficient

- the issue with the sampling distribution for Pearson's r is that the sampling distribution is not normal
  - to fix this, we can adjust the sampling distribution to be normal by using z-scores

$$z_r = \frac{1}{2} log_e(\frac{1+r}{1-r})$$

- the z_r also has a standard error by calculating the following equation

$$SE_{z_r} = \frac{1}{\sqrt{N-3}}$$

Significance of Correlation Coefficient

- transform your adjusted r value into a z-score
  - our hypotheses for a correlation is that the correlation will be different from zero
  - as with other tests, rather than subtract zero, we can just have the value divided by the standard error to get a z-score

$$z = \frac{z_r}{SEz_r}$$

- we could also use a t-test with a correction in the degrees of freedom $(N - 2$

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Some Notes About Correlation Coefficients

- remember that our correlation coefficients do not indicate causality
    - unless the design of your study would indicate a cause –> effect relationship (e.g., experiments), then you can only state that there is a relationship present
        - or state there is evidence of a significant relationship
- this can be due to several reasons
    - directionality of your variables
        - your IV could be your DV and vice versa
        - Ex: depression –> obesity | obesity –> depression
    - is there a third variable or **tertium quid** which could be influencing the relationship between your two variables

Confidence Intervals for r

- we can calculate confidence intervals using those z_r values and the corresponding standard errors

$$lower\ CI = \overline{X} - (1.96 * SE)$$

$$upper\ CI = \overline{X} + (1.96 * SE)$$

- becomes

$$lower\ CI = z_r - (1.96 * SE_{z_r})$$

$$upper\ CI = z_r + (1.96 * SE_{z_r})$$

Confidence Intervals for r

- we can then convert these back to correlation coefficients by using the following formula

$$r = \frac{\epsilon^{2z_r} - 1}{\epsilon^{2z_r} + 1}$$

- SPSS does not compute your standard confidence intervals for you
  - to get around this AGAIN we will be using bootstrapped confidence intervals

Bivariate Correlation

- let's talk about some sources of bias
- when fitting a linear model, we want a linear relationship between our variables so we need
  - an outcome that is numeric/continuous/ratio/interval
  - and a predictor that is also numeric/continuous/ratio/interval
- we'll also look for outliers
  - there are additional correlational tests that can rank the data to deal with outliers
  - JP: I don't think outliers will affect our data anyway
- we'll also look at out P-P and Q-Q plots to make sure the data looks normal

Pearson's Correlation Coefficient Using SPSS Statistics

- we will cover SPSS stuff during the activity section
  - if you square your correlation coefficient you get your **coefficient of determination**
  - which is the amount of variability in one variable that is shared by the other variable
  - R^2

Spearman's Correlation Coefficient

- **Spearman's correlation coefficient**, denoted as r_s, is a non-parametric statistic that can be useful for minimizing effects of extreme scores (outliers) or violations of assumptions
- rquires only ordinal data for both variables
  - pronounced as rho or ρ
- it works by ranking the data of your variables and then applies the Pearson's correlation coefficient equation to those ranks
- we will look at Spearman's correlation coefficient when conducting correlations in SPSS

Kendall's tau τ (non-parametric)

- non-parametric test that should be used for small datasets
  - with large number of "tied" ranks
- seen as a better alternative to Spearman's correlation as an estimate of the correlation in the population

Biserial & Point-Biserial Correlations

- Biserial and point-biserial correlation coefficients are similar in that they are correlations where one variable is dichotomous (2 categories)
  - the difference is that dichotomous variable is either discrete or continuous
- a **point-biserial correlation coefficient** is used when one variable is a discrete dichotomous variable (sex)
- **biserial correlation coefficient** is used when one variable is a continuous dichotomous variable (passing an exam $= 1$, failing an exam $= 0$)
  - a variable on a continuum would fall under a biserial correlation coefficient

$$point - biserial = r_{pb}$$

$$biserial = r_b$$