Statistical Analysis Using Structural Equation Models

EPsy 8266

Christopher David Desjardins

Research Methodology Consulting Center

2/5/19

Topics

- Multiple regression
- Correlations

Simple linear regression activity

Main takeaways from the SLR activity

- Importance of initial data analysis
- Relationship between the standardized regression coefficient and the correlation
- How we could just solve for our coefficients knowing just the mean, standard deviations, and covariance of the variables.

One of the three components of causation (Bollen, 1989)

- What is contained in our residuals?
- What is the correlation of the residuals from a simple linear regression model with the predictor?
  - This allows us to **isolate** the effect of scores on the cubes test on scores on the visual test.
- But this isn't perfect. Because the observed scores of the visual test, as well as being a function of scores the cubes test, are also effected by the residuals.
- This is what is referred to as **pseudo-isolation** - β_1 is the effect of cubes test on the visual test isolated from the residuals.

Bollen's three components to causation

- The effect of X on Y can be **isolated**.

  This is the point of randomized control designs: To isolate the effect of X on Y from all other potential causes.

  Do you think randomized control designs successful do this?
- There should be an **association** between X and Y.
- There must be a **direction**

  X causes Y or Y causes X
- We will return to causality throughout the semester.

Multiple Regression

Continuing with the Holzinger-Swineford data set, what if wanted to predict scores on the visual test (`visual`, Y) given scores on the cubes (`cubes`, X_1) test and scores on the paper form board test (`paper`, X_2) ? Then, we would have multiple regression problem.
We can write this model as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

and because we care about inference, make a normality assumption:

$$Y | X_1, X_2 \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma^2)$$

We can fit this model in `R`

```
mod.mlr <- lm(visual ~ cubes + paper, hs.data)
```

Visual test model

```
summary(mod.mlr)

##
## Call:
## lm(formula = visual ~ cubes + paper, data = hs.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0123  -4.1119   0.5749   4.1018  15.7652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5419     2.4093   4.376 1.68e-05 ***
## cubes         0.3317     0.0803   4.131 4.70e-05 ***
## paper         0.7727     0.1336   5.782 1.86e-08 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.363 on 298 degrees of freedom
## Multiple R-squared:  0.1804,Adjusted R-squared:  0.1749
## F-statistic: 32.79 on 2 and 298 DF,  p-value: 1.346e-13
```

These are the **unstandardized partial regression coefficients**.

Obtaining beta weights

- How can we obtain **beta weights**?

Obtaining beta weights

- The **standardized partial regression coefficients**.

Beta weights

```
mod.std <- lm(scale(visual) ~ -1 + scale(cubes) + scale(paper),
              hs.data)
summary(mod.std)

##
## Call:
## lm(formula = scale(visual) ~ -1 + scale(cubes) + scale(paper),
##     data = hs.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.99979 -0.58703  0.08208  0.58558  2.25070
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## scale(cubes)  0.22304    0.05391   4.138 4.57e-05 ***
## scale(paper)  0.31221    0.05391   5.792 1.76e-08 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9068 on 299 degrees of freedom
## Multiple R-squared:  0.1804,Adjusted R-squared:  0.1749
## F-statistic:  32.9 on 2 and 299 DF,  p-value: 1.219e-13
```

Calculating betas

Beta weights can be calculated if we know the correlations among these 3 variables.

To calculate the standardized partial regression weight of cubes on visual, we would do the following.

$$b_{x_1} = \frac{r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}}{1 - r^2_{x_1 x_2}}$$

This takes the correlation between visual and cubes and adjusts for the correlation between cubes and paper and paper and visual and divides by the total variance of visual with the shared variance between cubes and paper removed.

What is b_{x_1} equal to if there is no correlation between cubes and paper?

Calculating betas

$r_{x_1 y} = 0.297, r_{x_1 x_2} = 0.238, r_{x_2 y} = 0.365$

```
(0.297 - 0.238 * 0.365) / (1 - 0.238^2)

## [1] 0.2227473
```

How would we solve for b_{x_2}?

Coefficient of determination

The coefficient of determination, i.e., the proportion of variance in Y that is explained by our model, can be calculating using the beta weights and the sample correlations:

$$R^2 = b_{x_1} * r_{x_1 y} + b_{x_2} * r_{x_2 y}$$

It can be obtained in `R` as follows:

```
summary(mod.std)$r.square

## [1] 0.1803702
```

Review of regression - Assumptions

Assuming standardized regression coefficients

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Assumptions that must be true for OLS estimates to be **unbiased** for the true model parameters.

1. Y, X_1, and X_2 are measured without error (i.e. reliability equal to 1)
2. ϵ is independent of X_1 and X_2
3. The relationship specified is correct, that is, e.g. Y is linearly related to X_1 and X_2

Shoe size and vocab

The classic examine used to demonstrate the partial correlation and **spuriousness** is shoe size and vocabulary breadth.

Consider the following correlation matrix between shoe size, vocabulary breadth, and age:

shoe size (X)	1		
vocabulary breadth (Y)	.5	1	
age (W)	.8	.6	1

- Please interpret the correlations in this matrix.

Shoe size and vocab

The classic examine used to demonstrate the partial correlation and **spuriousness** is shoe size and vocabulary breadth.

Consider the following correlation matrix between shoe size, vocabulary breadth, and age:

shoe size (X)	1		
vocabulary breadth (Y)	.5	1	
age (W)	.8	.6	1

- What might happen if we calculate the partial correlations?

Partial correlation

$$r_{XY*W} = \frac{r_{XY} - r_{XW} r_{WY}}{\sqrt{(1 - r_{XW}^2)(1 - r_{WY}^2)}}$$

shoe size (X)	1		
vocabulary breadth (Y)	.5	1	
age (W)	.8	.6	1

Suppression

Now, suppose we had the following (example from Kline)

Amount of psychotherapy (X)	1		
Number of prior suicide attempts (Y)	.19	1	
Degree of depressions (W)	.49	.70	1

- Please interpret the correlations in this matrix.

Suppression

Now, suppose we had the following (example from Kline)

Amount of psychotherapy (X)	1		
Number of prior suicide attempts (Y)	.19	1	
Degree of depressions (W)	.49	.70	1

- Now calculate the relationship between X and Y partialling out W.

Part correlations

Sometimes we are interested in removing a third variable W from only one of the variables. (Part correlations)
The equation for partialing W out of X but not Y is shown below:

$$r_{Y(X*W)} = \frac{r_{XY} - r_{XW} r_{WY}}{\sqrt{(1 - r_{XW}^2)}}$$

Amount of psychotherapy (X)	1		
Number of prior suicide attempts (Y)	.19	1	
Degree of depressions (W)	.49	.70	1

- Calculate the part correlation between X and Y partialling W out of X only.

Part correlations

Sometimes we are interested in removing a third variable W from only one of the variables. (Part correlations)
The equation for partialing W out of X but not Y is shown below:

$$r_{Y(X*W)} = \frac{r_{XY} - r_{XW} r_{WY}}{\sqrt{(1 - r^2_{XW})}}$$

Amount of psychotherapy (X)	1		
Number of prior suicide attempts (Y)	.19	1	
Degree of depressions (W)	.49	.70	1

- Calculate the part correlation between X and Y partialling W out of X only.
- Is it larger or smaller than the partial correlation?

Tetrachoric correlation from Wirth & Edwards, 2007

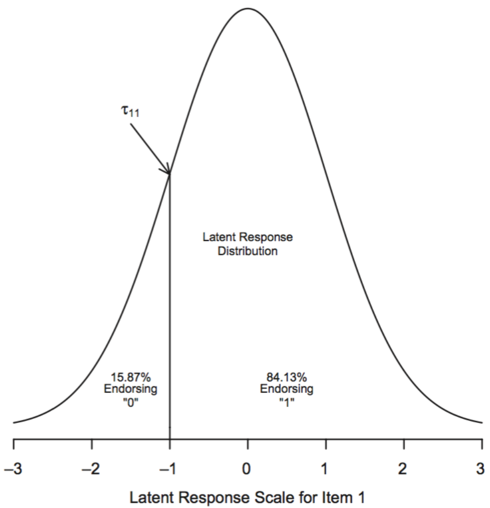

Figure 1. Latent response distribution for a single dichotomous item representing the latent distribution of interest. τ_{11} marks the latent cut-point between observed responses.

Tetrachoric correlation from Wirth & Edwards, 2007

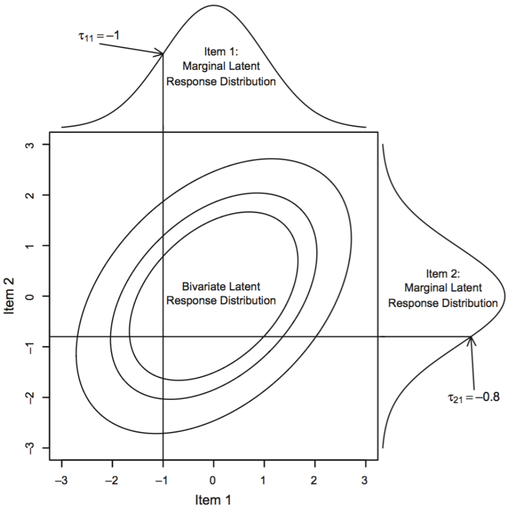

Figure 2. Bivariate and marginal latent response distributions for two dichotomous items. The bivariate latent response distribution, with a correlation of .70, represents the distribution of interest. The ellipses represent the .01, .05, and .10 regions. The threshold parameters τ_{11} and τ_{21} denote the cut-points for Items 1 and 2, respectively.